Reliability Generalization of three frequently used Open Access Measures: MAAS, SEK-27, and PSQ

Gülay Karadere[1], Nadine Wedderhoff[2], and Michael Bosnjak[1,2]
guek@leibniz-psychology.org, naw@leibniz-psychology.org, mb@leibniz-psychology.org

[1] Leibniz Institute for Psychology Information (ZPID), Trier, Germany
[2] University of Trier, Trier, Germany

Introduction

Background

The Leibniz Institute for Psychological Information and Documentation's (ZPID) Open Test Archive is a repository with approximately 200 Open Access tests from psychology and related disciplines. These are primarily research instruments whose psychometric quality has been sufficiently verified and documented. However, the quality information reported for these tests represents narrative test evaluations only, with their systematic evaluation still lacking. Within this project, for the first time, a selection of these test instruments will be analyzed using reliability generalization meta-analyses.

Reliability generalization (RG) is a meta-analytic method for empirically examining score reliability estimated for one measure or a group of measures based on the same construct. It was first introduced by Vacha-Haase in the late 1990s to (a) identify the typical score reliability for an instrument across all studies, (b) estimate the amount of variability in the score reliability across all studies, and (c) discover the influencing factors of the variability (Vacha-Haase, 1998; see also Vacha-Haase & Thompson, 2011).

The starting point of RG was the recognition that neither the statement "the reliability of the test" nor the statement "the test is reliable" (Thompson, 1994) correctly reflect what reliability expresses: A reliability coefficient (for internal consistency (e.g., Cronbach's alpha, K-R 21), retest reliability, parallel forms reliability, or split-half reliability) describes a score estimated in a particular sample. This score, calculated for a specific sample, cannot be transferred to another sample. Nevertheless, some researchers transfer or report score reliability from other sources. Others, however, do not even mention reliability. This practice is called reliability induction. Thompson and Vacha-Haase (2000) noted that reliability induction is problematic for researchers who conduct meta-analyses. In their examination of the treatment of score reliability they found that more than half of the primary reports did not mention score reliability (Vacha-Haase & Thompson, 2011). To show and emphasize the importance of determining reliability and providing reliability estimates in empirical research, they refer to the corresponding statement of the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force, 1999).

> “It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric” (Wilkinson & APA Task Force, 1999, p. 596).

With the fact that reliability is not a stable part of a test, each application of a test provides a score which may vary according to different samples and measurement conditions. Botella, Suero, and Gambara (2010) described three sources of variation (for Cronbach's alpha): (1) sampling scheme/variance, (2) error variance, and (3) correlation between the items. Due to different strategies of recruitment of the participants, different sampling schemes result from the population and cause different population variance estimates and, as a consequence, impact the score reliability. As the second source, the error variance varies depending on the intensity of a moderator's influence. The correlation between the items can also be influenced by moderators. Botella and colleagues (2010) reached the following conclusions:

1. The empirical variances and the reliability coefficients have an inverse relationship when the variability of the variances is caused by differences in the error variance.
    - ➢ Larger error variances lead to larger sample (empirical) variances and smaller reliability or internal consistency coefficients.
2. The empirical variances and the reliability coefficients have a direct relationship when the variability is caused by different sampling schemes or variations in the correlations between the items.
    - ➢ Larger alpha coefficients come along with a simultaneous increase in sample variances.

The aim of this work is to conduct RG studies with three selected open access test instruments. For this, the test selection was carried out with the tests from the Open Test Archive based on the following criteria:

<u>1. High usability: Examination of the test download numbers: Top 15</u>

The first step was to determine which tests exhibit high usability, that is, which tests are accessed most frequently by users. Information on this aspect can be found by examining the download numbers for each test. In addition, download numbers can provide an indication of its popularity. For the present investigation, the first 15 tests in the download ranking were selected for shortlisting.

2. High application: Quick search: Top 6

In addition to the high usability numbers, the application of these 15 tests has been examined in a quick search to ensure that they have been (often) applied and that a sufficient number of published reports are available. The search terms were the names of the tests. This search (via Google Scholar on December 16, 2019, 21:25 h) yielded hit numbers varying from 8 to 124,000. Those with approximately 100 hits as a realistic value for the search result and good data basis for the meta-analysis were shortlisted, so that it end up with six tests.

3. High citations: Citation analysis of the publications: Top 3

The third step focused on uncovering the number of citations of the original test publication for the six tests resulting from the quick search in Step 2. For this purpose, citation numbers were searched for in three academic databases: Google Scholar, Web of Science, and Scopus. The search terms were the reference source of the test authors in which the tests are mentioned for the first time. The three tests with the least number of citations in journal articles were eliminated.

The top three open access tests, determined by the three-part selection process described above, are listed in the following. Because two of these tests are German translations of original English measures, and when conducting broad RG studies, all (language) versions of a test are included, the original tests will be the anchor for this work and are listed here as well.

1) MAAS - The German version of the Mindful Attention and Awareness Scale (Michalak, Heidenreich, Ströhle, & Nachtigall, 2008);
   Original: Mindful Attention and Awareness Scale (MAAS; Brown & Ryan, 2003);
2) SEK-27 - Selbsteinschätzung emotionaler Kompetenzen-27 (Berking & Znoj, 2008);
   English version: Emotion Regulation Skills Questionnaire (ERSQ; Grant, Salsman, & Berking, 2018);
3) PSQ - The German version of the Perceived Stress Questionnaire (Fliege, Rose, Arck, Levenstein, & Klapp, 2001);
   Original: Perceived Stress Questionnaire (PSQ; Levenstein, Prantera, Varvo, Scribano, Berto, Luzi, & Andreoli, 1993).

Mindfulness Attention and Awareness Scale (MAAS; Brown & Ryan, 2003) is an instrument with 15 items measuring the ability to focus attention on the present moment and act with mindfulness. According to Kabat-Zinn (1990, 2003), mindfulness can be defined as (1) an intentional and (2) nonjudgmental form of attentional control, which happened (3) in the present moment. All items describe negative statements about frequent experiences of cognitive, emotional, physical, interpersonal, and general issues and are answered on a six-point Likert scale. High scores in MAAS represent a person with a high level of mindfulness.

Brown and Ryan (2003) examined the internal consistency (Cronbach's alpha) and the stability of MAAS after a period of four weeks. For both, they report a score of .81. Applying the MAAS in a sample of cancer patients and healthy subjects, Carlson and Brown (2005) achieved alpha scores of .87 in both samples. For the German version of the MAAS, an alpha coefficient .83 and a stability coefficient of rtt = .82 was found after 21 days (Michalak, Heidenreich, Ströhle, & Nachtigall, 2008).

On the basis of a factor analysis, unidimensionality of the MAAS could be confirmed, which explained 95% of the variance (Brown & Ryan, 2003, Carlson & Brown, 2005). The MAAS can differ between mindful and mindless persons and between meditators and nonmeditators. Brown and Ryan (2003) report correlations between high MAAS scores and self-consciousness and life satisfaction as well. Furthermore, correlations between the scales of the personality measures NEO-PI (Costa & McGrae, 1985) and NEO-FFI (Borkenau & Ostendorf, 1993), the 20-item version of the Beck Depression Inventory (BDI; Beckham & Leber, 1985) or the State-Trait Anxiety Inventory (STAI; Spielberger, 1983) could be found with the original and the German version of MAAS (Brown & Ryan, 2003; Michalak, Heidenreich, Ströhle et al., 2008).

Low scores are associated with depressiveness or anxiety (Brown & Ryan, 2003). Patients with cancer and depression show increased scores on the MAAS after a clinical treatment (Carlson & Brown, 2005; Michalak, Heidenreich, Meibert, & Schulte, 2008).

Emotion Regulation Skills Questionnaire (ERSQ) is the English name of the original German measure, the SEK-27 ("Selbsteinschätzung emotionaler Kompetenzen-27" [Self-report measure of emotional competencies-27-item version], which was developed by Berking and Znoj (2008). The regulation of emotions is essential for well-being. For a successful emotion regulation, the following nine skills based on the ACE model (adaptive coping with emotions model; Berking, 2008, 2010; Berking & Whitley, 2014) are needed: (1) Awareness, (2) Understanding, (3) Effective Self-support, (4) Readiness to Confront, (5) Acceptance, (6) Tolerance, (7) Modification, (8) Sensation, and (9) Clarity.

The ERSQ consists of 27 items to assess nine subscales that reflect the skill components of the ACE model. Responding to how they dealt with negative emotions in the last week, subjects indicate their agreement on a 5-point Likert scale ranging from 0 (not at all) to 4 ((almost) always). The scores for each subscale and the total scale indicate the degree of emotion regulation skills.

Although it has been translated into English for a broad application by Grant, Salsman, and Berking (2018), the first validation of the ERSQ was carried out on 576 healthy persons and a clinical sample in Germany (*n* = 238) (Berking & Znoj, 2008). The internal consistency (Cronbach's alpha) was between .68 and .81 for the subscales and .90 for the total scale. Grant and colleagues (2018) found higher coefficients for the English version. The retest reliability, with a time frame of two weeks, was between rtt = .48 and rtt = .74 for the subscales and rtt = .75 for the total scale. The three-week retest reliability of the English version again showed similar, low stability coefficients (Grant et al., 2018).

Exploratory and confirmatory factor analysis could confirm the nine-factor structure of the ERSQ. Based on cross-sectional and longitudinal investigations it is known that emotion-regulation skills have an impact on mental disorders (e.g., Levine, Marziali, & Hood, 1997; Seiffge-Krenke, 2000), the relationship of the ERSQ with well-being and mental health has been measured. An association between the scales and the indicators could be found. Furthermore, it can differ between healthy and clinical samples. The scores of the clinical sample increased after receiving a psychotherapeutic treatment and provided evidence for change sensitivity of the ERSQ (Berking & Znoj, 2008).

The ***Perceived Stress Questionnaire*** (PSQ; Levenstein et al., 1993) was simultaneously constructed as a 30-item version in English and Italian with seven subscales: (1) Worries, (2) Tension, (3) Lack of Joy, (4) Fatigue, (5) Harassment, (6) Overload, and (7) Irritability. It measures how individuals experience, evaluate, and cope with stress. The PSQ Index estimates, on a dimensional level from 0 (lowest possible level of stress) to 1 (highest possible level of stress), how strong the stress level is. A score can also be calculated for the general (past year or two) and current (past month) stress level.

The PSQ-30 was completed by outpatients, inpatients, students, and health care workers. The internal consistency (Cronbach's alpha) was between .90 and .92, and the retest-reliability after a period of 7 to 10 days was rtt = .82 to rtt = .86. For the examination of the construct validity, other stress relevant measures were assessed. Using factor analysis, seven factors could be extracted. The predictive validity could be found by comparing the PSQ scores, for example, of inpatients and outpatients, or of patients with an intestinal disease in a pre-post comparison regarding the degree of complaints.

The German version was developed by Fliege et al. (2001). Among other things, they examined the factorial structure of the measure, and as a consequence of low factor loadings of some items, it results with a 20-item version with four scales (worries, tension, lack of joy, demands).

Objectives

Reliability generalization allows researchers to aggregate reported reliability coefficients obtained from the application of a test, which is based on the same construct. That allows to examine the variability due to different test applications and study variables. Thus, score reliability is the focus of this research. For this purpose, the original versions of the three test measures fulfilling the criteria mentioned above (i.e., high usability, high application, and high citation) will be the object of investigation. Furthermore, all findings from the primary studies will be available in PsychOpen CAMA (Community Augmented Meta-Analyses; another ZPID service) for replicating and modifying meta-analyses. By adding subsequent findings the evidence will be keep up-to-date.

A typical RG study is designed to answer the following three questions:

1) What is the average score reliability coefficient across the studies?

2) How large is the systematic variability (heterogeneity) of the score reliability across all studies?

3) Which study characteristics influence the score reliability across the studies?

This work aims to answer these questions for MAAS, SEK-27, and PSQ. The main focus is on the examination of the moderator variables, as it is expected that the study characteristics are associated with and influence reliability.

Method

Selection Criteria

Specifying eligibility criteria "serve[s] to ensure that studies are selected in a systematic and unbiased manner" (Liberati et al., 2009, W-72). They guide the steps of a meta-analysis and make them comprehensible. For studies to be included, all selection criteria should be made explicit (e.g., Dieckmann, Malle, & Bodner, 2009). A variation in criteria "might lead to systematic differences in which studies remain in the synthesis" (Cooper, Hedges, & Valentine, 2019, p. 9). Therefore, each study should be selected on the basis of a previously established catalog of criteria which (a) correspond with the research question or objectives and (b) help to guide the next steps of meta-analysis as developing the search strategy. Hence, the selection criteria may determine the usefulness of relevant articles.

At the beginning of the RG study, eligibility criteria that guide the subsequent decisions on which studies should be included have to be defined. In the present case, the selection criteria for the studies include the following two criteria. First, all studies have to apply measures based on the same construct. This can be the implementation of any form of the three questionnaires MAAS, SEK-27/ERSQ, or PSQ. Second, they should report a reliability estimate. The selection here is restricted to the report of at least alpha and/or retest reliability coefficients estimated on an own sample.

There will be no limitation regarding the specific test version, as all modifications and adaptations of these measures are taken into consideration (e.g., short vs. long form, paper-pencil vs. online, with different rating scales, self vs. other assessment form as well as versions for different target groups will be included). The tests can be applied on different samples (nonclinical vs. clinical participants) and under different conditions for implementation. However, restrictions are defined regarding the language of the publication, as only German and English articles are considered.

Search Strategies

After specifying the selection criteria, a representative sample of articles and research objects on the research topic have to be found. There are different search strategies to achieve this. In general, researchers should implement a combination of several distinct search strategies, in order to minimize the risk of systematic differences between found and unfound studies (Cooper, 2017). Three "*channels*" are defined by Cooper (2017): (1) researcher-to-researcher channels, which include the personal contact and mass solicitation, (2) quality-controlled channels like conference papers or peer-reviewed journals, and (3) secondary channels. The latter includes reference lists, research bibliographies, prospective research registers, the internet, databases, and citation indexes. An adequate literature search according to Cooper (2017) should include a search in scientific databases, a review of relevant journals, backward and forward literature reference searching, and contacting active and known researchers. The latter is of particular importance for obtaining unpublished manuscripts. Avoiding publication bias and to have comprehensive study data, it is necessary to include publications of various types (including, e.g., gray literature). An appropriate literature search across several accesses can be used to include a representative set of studies, which allows a generalization of the research question (Cooper, 2017).

Every single search process will be recorded and illustrated in a flow chart diagram according to the *PRISMA model* (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009; Liberati et al., 2009). *Figure 1* helps to clarify how many studies were found in different channels (*Step 1: Identification*), how many were screened (*Step 2: Screening*), how many have been subjected to an in-depth full-text screening (*Step 3: Eligibility*), and how many remaining articles could be included (*Step 4: Included*) for the RG study. These steps are described in more detail below.

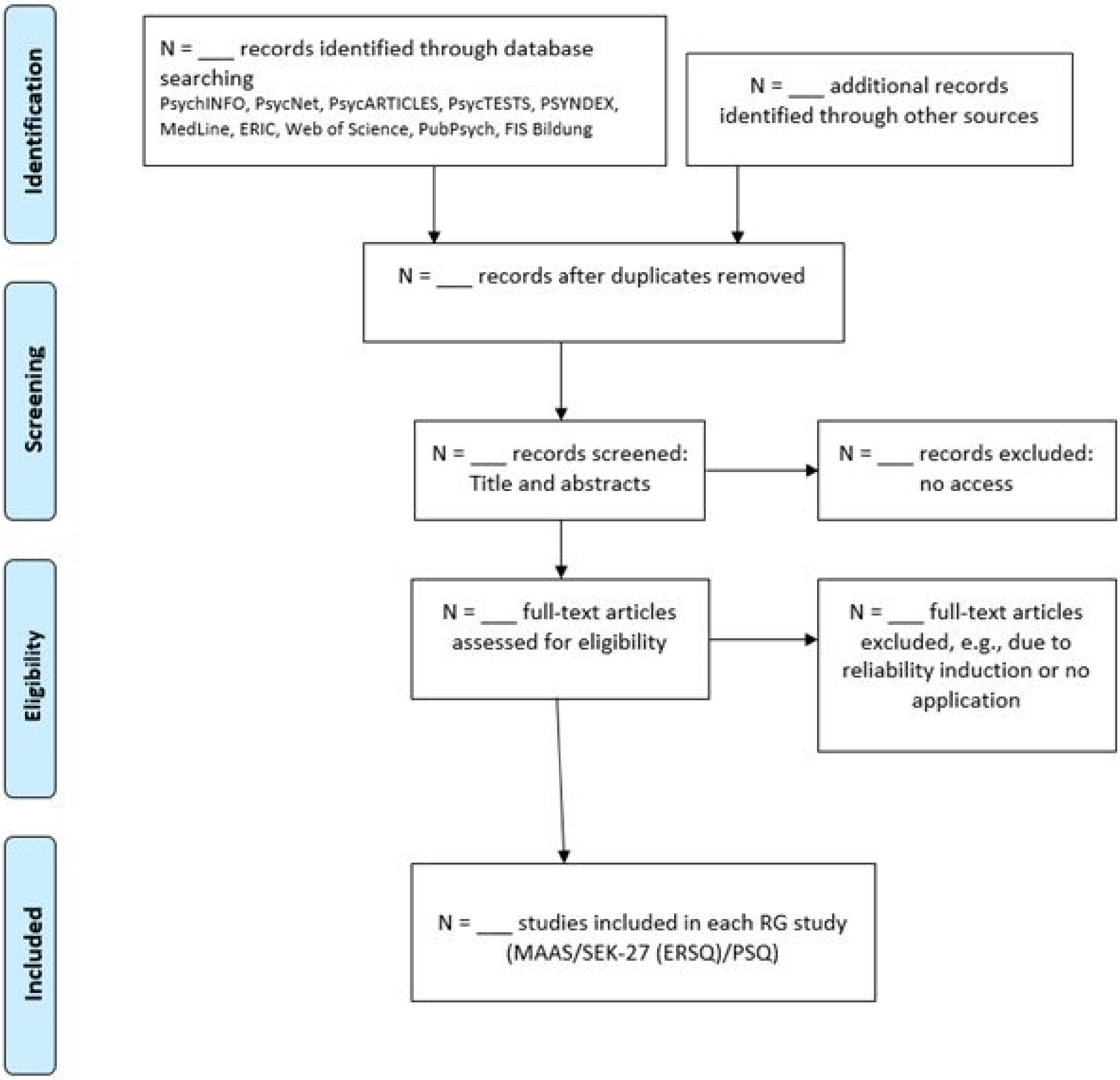
Identification
Screening
Eligibility
Included
N = ___ records identified through database searching
PsychINFO, PsycNet, PsycARTICLES, PsycTESTS, PSYNDEX, MedLine, ERIC, Web of Science, PubPsych, FIS Bildung
N = ___ additional records identified through other sources
N = ___ records after duplicates removed
N = ___ records screened: Title and abstracts
N = ___ records excluded: no access
N = ___ full-text articles assessed for eligibility
N = ___ full-text articles excluded, e.g., due to reliability induction or no application
N = ___ studies included in each RG study (MAAS/SEK-27 (ERSQ)/PSQ)

Figure 1. Flow chart of the search process.

Step 1: Identification

The first step is to locate relevant studies for alle three measures by searching several electronic databases: PsycINFO, PsycNet, PsycARTICLES, PsycTESTS, PSYNDEX, MedLine, ERIC, and Web of Science. PubPsych, as a search portal including several databases, allows additional reports to be located for all three selected tests. An internet search via Google Scholar will also identify additional studies.

FIS Bildung is the German database of the Leibniz Institute for Human Development (DIPF) and contains educational literature which should be particularly relevant for the MAAS because of its frequent application in this field. MedLine and Eric will be an important source of information for the PSQ, and possibly for the SEK-27 because of the relevance of emotions in the medical context.

The search terms will contain the original measure names in English and German—an exception is the original name of PSQ, which was not translated into German—as well as the respective abbreviation, and will be connected via Boolean operators: [“MAAS” OR “Mindful Attention Awareness Scale” OR “Achtsamkeits-Aufmerksamkeits-Bewusstseins-Skala”], [“ERSQ” OR “Emotion Regulation Skills Questionnaire” OR “SEK-27” OR “Selbsteinschätzung emotionaler Kompetenzen-27”], and [“PSQ” OR “Perceived Stress Questionnaire”]. The search time period depends on the first publication of each measure. Accordingly, the literature searches for the PSQ, the MAAS, and the SEK-27 encompass articles published since 1993, 2003, and 2008, respectively.

In addition, journals, which have a strong focus on the constructs mindfulness (e.g., *Mindfulness*), stress (e.g., *Stress*, *Stress and Health*), and emotion (e.g., *Emotion, Cognition & Emotion*) or assessment (e.g., *European Journal of Psychological Assessment*) will be reviewed, a backward and forward literature reference search will be conducted, and a review of previous conference papers will be also conducted for the purpose of finding eligible studies. As a final search procedure, several authors will be contacted via email, including those who constructed or frequently applied these measures.

Step 2: Screening

At the beginning of the screening process, all identified duplicates will be removed. Subsequently, all remaining articles will be screened to see whether the measures or the targeted constructs are mentioned in the title or in the abstract. If so, the respective articles are reviewed in-depth, and these might be categorized as eligible, whereas all others with no mention of the targeted constructs might be excluded.

Step 3: Eligibility

All studies mentioning the test in the title and/or abstract will be reviewed to ensure that (1) the MAAS, SEK-27, and PSQ are implemented and (2) they report a reliability estimate computed on their own sample rather than (a) taken from a previous study sample or from another source like a test manual or another application of the measure (*reliability induction by report*) or (b) a reliability coefficient is not even reported (*reliability induction by omission;* Shields & Caruso, 2004; Thompson, 2003). Thompson and Vacha-Haase (2000) noted that reliability induction poses a potential problem when low score reliabilities have a negative impact on subsequent data analyses. The main problem is rather the resulting reduced amount of reported score reliabilities.

Studies that do not apply one of the measures and those that do apply one of the three measures but have induced reliability by omission, will be excluded. In the case of reported reliability induction, an effort will be made to find the original study and use the original reliability information for the RG study (Whittington, 1998). Another option is the calculation of the reliability estimate through the Kuder-Richardson formula 21 (K-R 21; Kuder & Richardson, 1937) for dichotomous items.

In the end, only the studies of the not induced score reliability will remain.

Step 4: Included

At the end of the search process, the number of included studies and the reliability estimates are reported.

Data Extraction

In general, there are three major aims when conducting an RG: (1) To obtain an average reliability estimate across studies, (2) to assess the amount of heterogeneity between effect size coefficients (i.e., reliability estimates), and (3) to evaluate the potential moderating effect of study and sample characteristics on measurement accuracy (Henson & Thompson, 2002, Rodriguez & Maeda, 2006). Therefore, for every eligible study, alpha and retest coefficients, the respective sample variance, as well as several study and sample characteristics as moderator variables will be extracted. As sources of variability in score reliability (Vacha-Haase, 1998), they will be needed for conducting a moderator analysis and are part of the coding process. For this procedure, a coding scheme must be prepared and the collected data should ideally be entered by two independent coders. The following checklist of moderator variables from the REGEMA model (*reliability generalization meta-analyses*; Sánchez-Meca et al., 2019, see also Sánchez-Meca et al., 2011) provides guidance for the coding: (a) sample size, (b) mean and standard deviation of the total test scores, (c) mean and standard deviation of the subscales (MAAS: no subscales; SEK-27: 9 subscales; PSQ-30: 7 subscales, PSQ-20: 4 subscales), (d) original test version versus other, (e) test length, (f) mean and standard deviation of the age of participants (in years), (g) gender distribution (in percentages), (h) country/culture, (i) target population, (j) disorder of the participants (for PSQ and SEK-27), (k) mean and standard deviation of disorder history (in years), (l) study focus (psychometric vs. substantive), (m) focus of the psychometric studies (MAAS, SEK-27, PSQ vs. other tests), (n) researcher affiliation, (o) publication year of the study, (p) instruction, and (q) interval between two measurements (for retest reliability). The last two variables are not part of the REGEMA model. For a better overview and clear separation by test, all these moderator variables are shown in Table 1.

Table 1
Relevant Study Characteristics for MAAS, SEK-27, and PSQ with Moderating Effect

	Measures		
	MAAS	**SEK-27**	**PSQ**
Study characteristics	● mean of the total test scores a) ● standard deviation of the total test scores a) ● mean of the age of participants (in years) a) ● standard deviation of the age of participants (in years) a) ● gender distribution (in percentages) a) ● original test version vs. other b) ● disorder of the participants a) ● mean of disorder history (in years) a) ● standard deviation of disorder history (in years) a) ● study focus b) ● focus of the psychometric studies b) ● researcher affiliation b) ● publication year of the study a) ● target population b) ● country/culture b) ● language of the test b) ● instruction b) ● interval between two measurements a)		
		● mean of the subscales a) ● standard deviation of the subscales a)	
			● test length a)

Note. Moderators: a) continuous vs. b) categorical.

Each study will be reviewed to determine which reliability coefficient is reported. From previous RG studies, we know that one or two scores are typically calculated. The most common one is Cronbach's alpha. Alpha estimates the internal consistency. It indicates the extent to which the single items of a test correlate with each other (Cronbach, 1951). Alpha is widely used because of its ability to provide a reliability estimate from a single measure implemented on a sample at a single time (see Greco, O'Boyle, Cockburn, & Yuan, 2018; Rodriguez & Maeda, 2006). Thus, the alpha coefficient will be examined for each RG study. A separate RG meta-analysis for the reported retest reliability will be included if there is sufficient data. The combination of different reliability coefficients in an RG study is not allowed due to different sources of error (Rodriguez & Maeda, 2006) .

To check the reliability of the coding process, the coding sheets of the two independent coders will be analyzed in terms of their inter-rater agreement. Potential inconsistencies should be solved by discussion to find consensus or by recommendations of a third person.

Statistical Analyses

In the following sections, the step-by-step procedures for the statistical analysis will be outlined in general. The first section (*Statistical Model*) describes the existing statistical models and details the model that is appropriate for this work. Section 2, *Pooling Reliability Estimates,* deals with the handling of all coefficients (e.g., [back]transformation, weighting). As the next step, the following section (*Heterogeneity Assessment*) describes the analysis options to determine whether the reliability coefficients are homogenous or not. In the fourth section (*Moderator Analysis*), the analysis of possible moderator variables will be presented.

The last two sections (*Assessment of Publication Bias & Sensitivity Analysis*) describe the risks of publication bias and the sensitivity analysis.

All analyses will be conducted with the statistical software R using the metafor package (Viechtbauer, 2010). Figure 2 depicts all the steps of the statistical analysis, including the functions of *metafor* for conducting the meta-analyses.

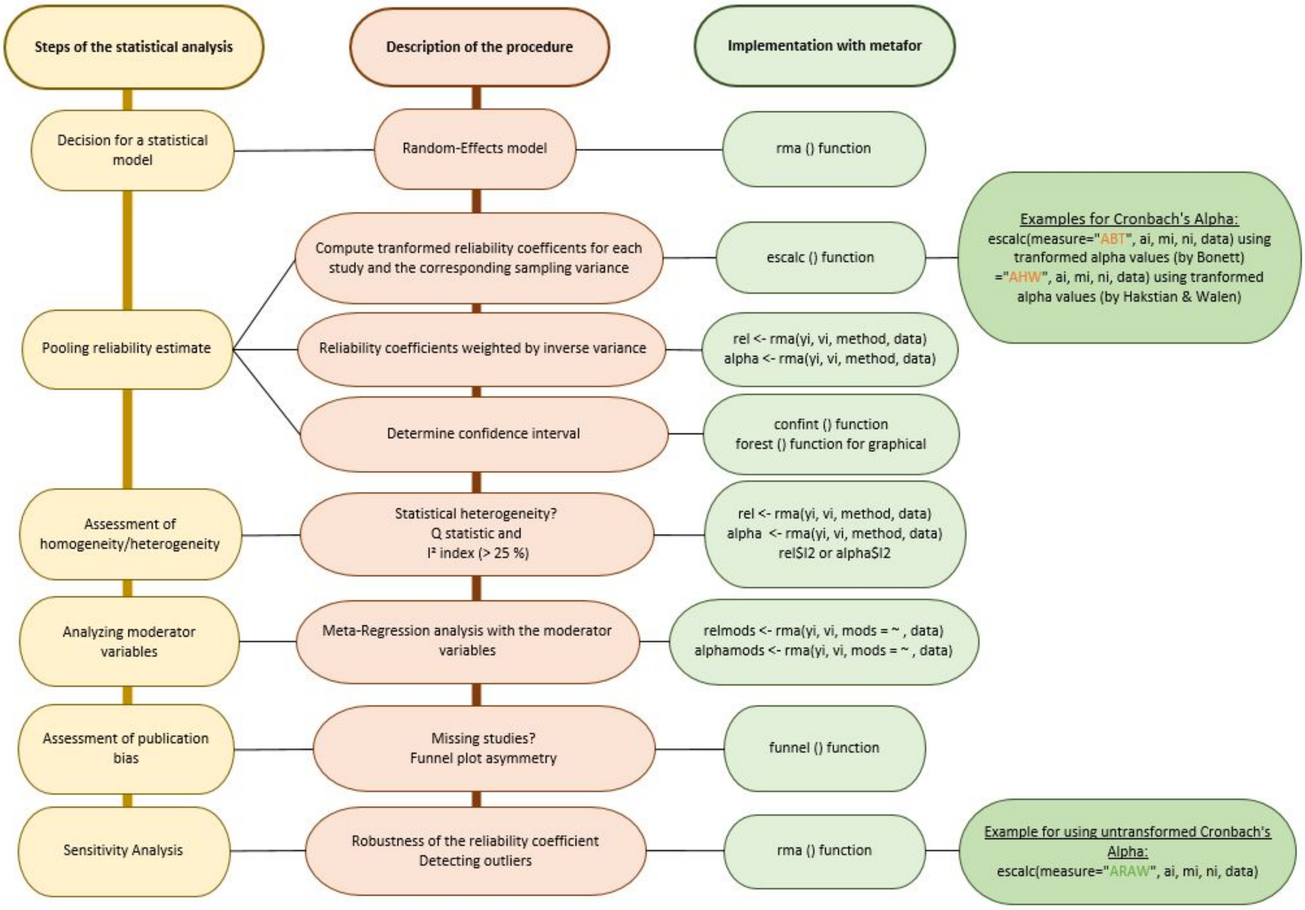
Steps of the statistical analysis
Description of the procedure
Implementation with metafor
Decision for a statistical model
Random-Effects model
rma () function
Compute tranformed reliability coefficents for each study and the corresponding sampling variance
escalc () function
Examples for Cronbach's Alpha:
escalc(measure="ABT", ai, mi, ni, data) using tranformed alpha values (by Bonett)
="AHW", ai, mi, ni, data) using tranformed alpha values (by Hakstian & Walen)
Pooling reliability estimate
Reliability coefficients weighted by inverse variance
rel <- rma(yi, vi, method, data)
alpha <- rma(yi, vi, method, data)
Determine confidence interval
confint () function
forest () function for graphical
Assessment of homogeneity/heterogeneity
Statistical heterogeneity?
Q statistic and
I² index (> 25 %)
rel <- rma(yi, vi, method, data)
alpha <- rma(yi, vi, method, data)
rel$I2 or alpha$I2
Analyzing moderator variables
Meta-Regression analysis with the moderator variables
relmods <- rma(yi, vi, mods = ~ , data)
alphamods <- rma(yi, vi, mods = ~ , data)
Assessment of publication bias
Missing studies?
Funnel plot asymmetry
funnel () function
Sensitivity Analysis
Robustness of the reliability coefficient
Detecting outliers
rma () function
Example for using untransformed Cronbach's Alpha:
escalc(measure="ARAW", ai, mi, ni, data)

Figure 2. Flow chart of the statistical analyses.

Statistical Model

The purpose of this section is to decide which statistical model is the appropriate one for calculating the variability in the reliability coefficients averaged across the included studies. Several approaches, like the fixed-effect (FE) model or the random-effects (RE) model, have gained acceptance (Cooper, 2017; Hedges & Vevea, 1998). The FE model "assumes that the samples of participants of the studies integrated are identical in composition and variability, and that the purpose of the meta-analyst is to generalize the results to a population of studies with identical characteristics to those included in the RG study" (Sánchez-Meca, López-López, & López-Pina, 2013, p. 408). In other words, it is expected that there is one true reliability estimate (effect size), and a variation in coefficients could only be caused by sampling variance (see also Cooper, 2017). The RE model makes the assumption that all included studies "have been randomly selected from a clearly defined superpopulation of potential studies, and then random samples of individuals are selected in each study to calculate a reliability coefficient" (Sánchez-Meca et al., 2013, p. 409). Thereby, all possible influences (study characteristics) are taken into account. The FE model considers only the variance due to the sampling of participants, whereas the RE model represents all the variance caused by the study characteristics of different applications of the same measure. Hence, it is also called the heterogeneity variance (Sánchez-Meca & Marín-Martínez, 2008). Based on these considerations, and with the fact that the assumption of the FE model is not valid for almost all data from practice (e.g., Hunter & Schmidt, 2000), the RE model is chosen as the underlying statistical model in these analyses.

Pooling Reliability Estimates

Before beginning the statistical analyses of RG, the reliability coefficient(s) for each measure has to be determined. The coding procedure will uncover the number of coefficients available in relation to the number of studies (to check, whether they are computed by different samples) and the specific reliability types. Following Rodriguez and Maeda's (2006) suggestion that "with sufficient numbers of each type of reliability value, each type of reliability should be synthesized separately," all the studies reporting the same type of reliability estimates will be merged and averaged separately (p. 309).

Regarding the calculation of an average reliability coefficient, different methods exist and reviewing these makes it clear that various decisions have to be made. The first decision affects the question whether reliability coefficients should be transformed before pooling. Although, there is some dispute on whether transformation is necessary or not (e.g., Henson & Thompson, 2002; Hunter & Schmidt, 2004), many researchers advocate a transformation to avoid an expected skewed distribution, for example, of alpha (e.g., Rodriguez & Maeda, 2006). The most common transformation method for correlations is the Fisher z-transformation (Borenstein, Hedges, Higgins, & Rothstein, 2009). Hence, test-retest and split-half reliabilities can be transformed to Fisher's z scores (Rodriguez & Maeda, 2006). Hakstian and Whalen's (1976) transformation method normalizes the distribution of the reliability coefficient. For Cronbach's alpha, a transformation by means of the Hakistan-Whalen formula should be preferred (e.g., Sánchez-Meca et al., 2013). As an alternative transformation method for alpha, Bonett (2002) developed a formula that also stabilizes the variances of alpha. To get accurate estimates, all combined reliability scores should be obtained from a large sample of independent studies (Bonett, 2010). This would result in narrower confidence intervals. That leads to the next step: the calculation of a confidence interval around the set of independent reliability coefficients for estimating the parametric effect size μ (Sánchez-Meca & Marín-Martínez, 2008). Bonett (2010) emphasized that a conventional RE confidence interval should not be used for pooling alpha reliability across multiple studies. The assumption of the RE model that studies are selected randomly from a superpopulation, which is normally distributed, is only an ideal case. A typical case of a meta-analysis contains either studies that are not randomly selected from a superpopulation with normal distribution or are randomly selected studies from a superpopulation without normal distribution.

Another aspect of pooling reliability coefficients is the weighting factor: The coefficients are commonly weighted by the inverse variance (Botella et al., 2010). It represents the amount of the within-study and the between-studies variances assumed by the RE model. The sampling variance of the transformed reliability coefficient(s) will be estimated with the corresponding formula.

The transformation value of the reliability coefficients and their confidence limits will be back-transformed to show the results in the original metric. It facilitates the conclusions of the findings (see Sánchez-Meca et al., 2013). To get an overview of the distribution of the reliability coefficients of the included studies, they are usually listed in a table. The estimates for the number of subscales of tests (for SEK-27/ERSQ and PSQ) can be illustrated in boxplots as graphical illustration.

Heterogeneity Assessment

The random-effects model assumes that the average reliability coefficient, as the true effect size, varies across the studies (Borenstein et al., 2009). This variation will be assessed by the *Q* statistic - automatically with the rma() function in metafor (Hedges & Olkin, 1985, Viechtbauer, 2010), and the I^2 index (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003). The former allows us to test whether the average coefficient is homogeneous, but the extent of an existing heterogeneity cannot be reported by the means of the *Q* test. Higgins and Green (2011) noted that "the test for heterogeneity is irrelevant to the choice of analysis; heterogeneity will always exist whether or not we happen to be able to detect it using a statistical test". Therefore, the I^2 index, as an addition to the *Q* test will be used. It assesses the degree of heterogeneity in percentages with values around 25% (= low heterogeneity), 50% (= medium heterogeneity), and 75% (= high heterogeneity) (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006, p. 198).

In general, a table will present a summary overview of all relevant results as the number of reliability coefficient, minimum and maximum of reliability coefficients, the unweighted and weighted mean, CI and the heterogeneity results obtained with the *Q* statistics and the I^2 indices separately for all test versions included.

Moderator Analysis

A moderator analysis will be conducted when the *Q* test is significant and the I^2 index indicates at least low heterogeneity (see above). In case of heterogeneity of the reliability coefficients, a weighted mean coefficient is not representative of the set of reliability estimates and it indicates that a moderator analysis should be carried out. Therefore, the influence of all study characteristics previously coded (see Table 1) will be examined by means of regression analyses. A meta-regression analysis will be conducted for the continuous and categorical moderator variables. The moderator variables will be used as predictors to explain at least part of the expected variability. To test the statistical significance of the moderator variables simultaneously or sequentially, the random effects meta-regression model is the best choice (Knapp & Hartung, 2003; see also Cooper, 2017; Sánchez-Meca et al., 2013). A significant fit allows an explanation of the influence of the moderator variables. They can be positively or negatively associated with the reliability, or can explain an increase or a decrease of the reliability (Botella et al., 2010).

The results of the moderator analysis can be presented in tables. A statistically significant relationship between a moderator variable and the reliability estimate can be shown in graphic form (e.g., scatter plot).

Assessment of Publication Bias

The phenomenon of publication bias is caused by missing studies, which remain unpublished or are difficult to access. However, information on reliability does not have a high influence on publishing a study. Another reason, particularly in RG studies, is reliability induction, whereby the amount of reliability coefficients reported from own samples is decreasing.

To analyze whether the intended RG studies failed to include a representative number of studies, a graphical representation by means of funnel plots will be conducted (Rothstein, Sutton, & Borenstein, 2005). An asymmetry of the funnel plot may indicate publication bias.

Sensitivity Analysis

Sensitivity analysis of the transformed and untransformed reliability coefficients will be conducted to estimate the robustness of the reliability coefficient. Additionally, the potential impact of outliers will be assessed. Outliers are observations outside of the majority of the data (Langford & Lewis, 1998). They can be caused by typing errors in the coding sheet or the primary researcher has submitted erroneous data, or the cause of outliers are unknown (see Cooper, 2017). One way to detect mild and extreme outliers is to use boxplots. Viechtbauer and Cheung (2010) recommend further approaches (e.g., studentized deleted residuals, Cook's distances) that can be applied in *metafor*. They examine the relation between the residuals and their corresponding standard error to find outliers.

Time plan

Milestone		Time range		
		MAAS	SEK-27	PSQ
Searching the literature	Enter potentially relevant studies into Mendeley	Jul 2020	Apr 2021	Dec 2021
Selecting the literature	Information about the eligibility for every study / or reasons of exclusion	Aug 2020	May 2021	Jan 2022
Develop Coding Scheme	Coding Scheme containing all variables of interest	01.-15. Sep 2020 (For the modification of the scheme for the other two tests an additional few days are needed later.)		
Coding	Codings of all included studies (including a second coder to calculate inter-rater reliability)	16. Sep - Nov. 2020	Jun - 15. Aug 2021	Feb - 15. Apr 2022
Analysis	Writing a reproducible code in R	Dez. 2020 (Later modifications for the other two tests can take some more days.)		
Report of the results	Writing the paper for the RG study	Jan - Mar 2021	~20. Aug - Nov 2021	~20. Apr - Jul 2022
Remaining part of the dissertation	Writing an abstract, an overall introduction and discussion part; reference list and appendix	Aug 2022 - Jan 2023		

References

Beckham, E. E., & Leber, W. (Eds.). (1985). *Handbook of depression: Treatment, assessment, and research.* Homewood, IL: Dorsey.

Berking, M. (2008). Training emotionaler Kompetenzen. Heidelberg: Springer.

Berking, M. (2010). *Training emotionaler Kompetenzen (2. Aufl.).* Heidelberg: Springer.

Berking, M., & Whitley, B. (2014). The adaptive coping with emotions model (ACE model). In *Affect regulation training* (pp. 19-29). New York, NY: Springer.

Berking, M., & Znoj, H. (2008). Entwicklung und Validierung eines Fragebogens zur standardisierten Selbsteinschätzung emotionaler Kompetenzen (SEK-27). *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 56*(2), 141-153. https://doi.org/10.1024/1661-4747.56.2.141

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

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological methods, 15*(4), 368.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to Meta-analysis*. Chichester: John Wiley & Sons, Ltd.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren- Inventar (NEO-FFI) nach Costa und McCrae [NEO-Five-Factor-Inventory according to Costa and McCrae]*. Göttingen, Germany: Hogrefe.

Botella, J., Suero, M., & Gambara, H. (2010, September 20). Psychometric Inferences From a Meta-Analysis of Reliability and Internal Consistency Coefficients. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/a0019626

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology, 84*(4), 822-848. https://doi.org/10.1037/0022-3514.84.4.822

Carlson, L. E. & Brown, K. W. (2005). Validation of the Mindfulness Attention Awareness Scale in a cancer population. *Journal of Psychosomatic Research, 58*, 29-33.

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach (Vol. 2)*. Sage publications.

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

Costa, P. T., Jr. & McGrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. https://doi.org/10.1007/BF02310555

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology, 13*(2), 101-115. https://doi.org/10.1037/a0015107

Fliege, H., Rose, M., Arck, P., Levenstein, S., & Klapp, B. F. (2001). Validierung des "perceived stress questionnaire" (PSQ) an einer deutschen Stichprobe [Validation of the "Perceived Stress Questionnaire" (PSQ) in a German sample]. *Diagnostica, 47*(3), 142-152. https://doi.org/10.1026//0012-1924.47.3.142

Grant, M., Salsman, N. L., & Berking, M. (2018). The assessment of successful emotion regulation skills use: Development and validation of an English version of the Emotion Regulation Skills Questionnaire. *PLoS ONE 13*(10): e0205095. https://doi.org/10.1371/journal.pone.0205095

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta‐analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies, 55*(4), 583-618. https://doi.org/10.1111/joms.12328

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*(2), 219-231. https://doi.org/10.1007/BF02291840

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and Random-Effects Models in Meta-Analysis. *Psychological methods, 3*(4), 486-504.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*(2), 113-127. https://doi.org/10.1080/07481756.2002.12069054

Higgins, J. P. T., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated May 2020]. The Cochrane Collaboration. Available from https://handbook-5-1.cochrane.org/

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta‐analysis. *Statistics in medicine, 21*(11), 1539-1558.

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

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

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta‐analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*(4), 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage publications.

Kabat-Zinn, J. (1990). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain and illness.* New York: Delacorte.

Kabat-Zinn, J. (2003). Mindfulness-Based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice, 10*(2), 144-156.

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

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. https://doi.org/10.1007/BF02288391

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 161*(2), 121-160.

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., & Andreoli, A. (1993). Development of the Perceived Stress Questionnaire: a new tool for psychosomatic research. *Journal of psychosomatic research*, *37*(1), 19-32. https://doi.org/10.1016/0022-3999(93)90120-5

Levine, D., Marziali, E., & Hood, J. (1997). Emotion processing in borderline personality disorders. *The Journal of nervous and mental disease, 185*(4), 240-246.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine, 151*(4), W-65-94. https://doi.org/10.1371/journal.pmed.1000100

Michalak, J., Heidenreich, T., Meibert, P. & Schulte, D. (2008). Mindfulness predicts relapse / recurrence in major depressive disorder following MBCT. *Journal of Nervous and Mental Disease, 196*(8), 630-633. https://doi.org/10.1097/NMD.0b013e31817d0546

Michalak, J., Heidenreich, T., Ströhle, G., & Nachtigall, C. (2008). Die deutsche Version der Mindful Attention and Awareness Scale (MAAS) psychometrische Befunde zu einem Achtsamkeitsfragebogen. *Zeitschrift für klinische Psychologie und Psychotherapie, 37*(3), 200-208. https://doi.org/10.1026/1616-3443.37.3.200

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International Journal of Surgery, 8*(5), 336-341. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological methods, 11*(3), 306-322. https://doi.org/10.1037/1082-989X.11.3.306

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester: John Wiley.

Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marín-Martínez, F., Rosa-Alcázar, A. I., & Gómez-Conesa, A. (2011). The Maudsley obsessive-compulsive inventory: a reliability generalization meta-analysis. *International Journal of Psychology and Psychological Therapy, 11*(3), 473-493.

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

Sánchez-Meca, J., López-Pina, J. A., Rubio-Aparicio, M., Marín-Martínez, F., Núñez-Núñez, R. M., López-García, J. J., & López-López, J. A. (2019, May 30). *REGEMA: Guidelines for Conducting and Reporting Reliability Generalization Meta-analyses*. ZPID (Leibniz Institute for Psychology Information). https://doi.org/10.23668/psycharchives.2476

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

Seiffge-Krenke, I. (2000). Causal links between stressful events, coping style, and adolescent symptomatology. *Journal of adolescence, 23*(6), 675-691. https://doi.org/10.1006/jado.2000.0352

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

Spielberger, C. D. (1983). *Manual for the State–Trait Anxiety Inventory: STAI (Form Y).* Palo Alto, CA: Consulting Psychologists Press.

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

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32. https://doi.org/10.3102/0013189X031003025

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Sage publications.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*(2), 174-195. https://doi.org/10.1177/0013164400602002

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

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*(3), 159-168. https://doi.org/10.1177%2F0748175611409845

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. http://www.jstatsoft.org/v36/i03/.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta‐analysis. *Research synthesis methods, 1*(2), 112-125. https://doi.org/10.1002/jrsm.11

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, *58*(1), 21-37. https://doi.org/10.1177/0013164498058001003

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. https://doi.org/10.1037//0003066X.54.8.594