# Utilizing Primary Study Quality in Meta-Analyses: A Step-by-Step Tutorial

Ronny Scherer[1,2] and Valentin Emslander[3]

[1] Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo

[2] Centre for Research on Equality in Education (CREATE), Faculty of Educational Sciences, University of Oslo

[3] Luxembourg Centre for Educational Testing (LUCET), Faculty of Humanities, Education and Social Sciences, University of Luxembourg

## Author Note

Ronny Scherer https://orcid.org/0000-0003-3630-0710

Valentin Emslander https://orcid.org/0000-0003-3690-5883

This study was pre-registered (https://doi.org/10.17605/OSF.IO/3E64F), and the respective data are openly available via OSF at https://doi.org/10.17605/OSF.IO/NGVCZ. We have no known conflict of interest to disclose. This work is part of the Centres of Excellence scheme, funded by the Research Council of Norway, project number 331640. This research was also supported in part by a grant from the Doctoral School in Humanities and Social Sciences of the University of Luxembourg to Valentin Emslander.

Correspondence concerning this article should be addressed to Ronny Scherer, Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Postbox 1161 Forskningsparken, NO-0318 Oslo. Email: ronny.scherer@cemo.uio.no.

This manuscript is a preprint. Status date: 16 June 2023.

## Abstract

Evaluating the quality of primary studies is a key step in meta-analytic reviews to reduce the risk of bias and establish the validity of the meta-analytic inferences. However, the extant body of research offers little guidance on how to represent and incorporate primary study quality (PSQ) in meta-analyses, and some common procedures, such as creating sum scores from a set of quality indicators, often lack the backing from measurement models. Addressing these issues, we present a tutorial that guides meta-analysts in their analytic decisions and approaches to represent and incorporate PSQ. Specifically, we describe, review, and illustrate approaches to (a) select or create quality indicators or scores a priori or as part of the meta-analytic model; (b) examine the possible moderator effects of PSQ; and (c) test the sensitivity of moderator effects to PSQ. We illustrate these approaches with three examples and present a step-by-step tutorial with analytic code for researchers' guidance. Overall, we argue for representing PSQ model-based if multiple quality indicators are available, the testing of moderator effects of PSQ on the effect sizes and their heterogeneity, and performing moderator sensitivity analyses.

*Keywords:* Primary study quality (PSQ), quality assessment, risk of bias, measurement models, sensitivity analysis

# Utilizing Primary Study Quality in Meta-Analyses: A Step-by-Step Tutorial

Evaluating and critically appraising the quality of primary studies has become a key step in meta-analytic reviews.[1-3] While primary study quality (PSQ) is defined differently across scientific disciplines and research contexts, the concept entails the extent to which a study's design, implementation, and purposes match.[4] Hence, evaluating PSQ in a meta-analysis contributes to crafting a validity argument for the reported effect sizes and provides evidence for or against the potential risk of bias.[5] Moreover, meta-analysts who critically appraise PSQ will obtain evidence on the credibility of the primary study and the meta-analytic effects and be able to contextualize seemingly suspicious findings in light of the PSQ.[6]

The extant body of research on PSQ in meta-analyses abounds in a plethora of quality definitions, assessments, scales, and checklists. This diversity has created several challenges: First, the ways in which meta-analysts have represented PSQ vary considerably and include but are not limited to single categorical indicators or scores, aggregates scores, and multiple yet separate indicators or scores.[7] For instance, using the same set of quality indicators, two meta-analysts may represent PSQ differently—while one meta-analyst aggregates these indicators to a total sum score, another one keeps them separate to represent the different dimensions of PSQ. This example illustrates the lack of methodological guidance that clarifies the key methodological decisions associated with the representation of PSQ in meta-analyses.[7]

Second, aggregating multiple quality indicators into PSQ scores often lacks the backing from the underlying statistical assumptions and models. For instance, creating a total sum score out of several quality indicators in a checklist relies on the assumptions that the set of indicators is unidimensional, and each indicator contributes to the overall score to the same extent.[8] However, these assumptions are often not met[9], as PSQ checklists often contain indicators of several aspects of PSQ and are thus likely multidimensional.[10] Hence, there is a clear need for crafting a validity argument for the use of aggregated PSQ scores in meta-analyses.[11]

Third, the strategies with which meta-analysts examine the impact of PSQ on the meta-analytic results are, by and large, limited to standard moderator analyses, yet without exploiting the full potential of these analyses.[11,12] For instance, most meta-analyses in psychology report on PSQ descriptively and present the linear moderator effects on the effect sizes.[4] While this approach allows meta-analysts to explore directly the relation between PSQ and the effect sizes, it has several caveats: This relation is assumed to be linear, although it may naturally be non-linear.[1] Moreover, moderator effects of PSQ often describe only the relation to the effect sizes. However, relations with PSQ may also exist with the heterogeneity between studies[13,14] or the effects of other moderators.[12] Extending the range of possible moderator effects, meta-analysts could gain more information about the sensitivity of the meta-analytic results to PSQ.[12]

In this paper, we address these challenges by presenting a methodological tutorial that guides researchers in their analytic decisions associated with representing and incorporating PSQ. These decisions include but are not limited to the a priori vs. meta-analytic selection or creation of quality indicators or scores, single vs. multiple quality indicators, aggregation of multiple quality indicators vs. keeping them separate, model-based vs. model-free quality score creation, moderator vs. moderator sensitivity analyses. We describe, review, and illustrate approaches to (a) select or create quality indicators and scores a priori or as part of the meta-analytic model; (b) examine the possible impact of PSQ on meta-analytic effect sizes and their heterogeneity; and (c) test the sensitivity of other moderator effects to PSQ. To support meta-analysts with hands-on guidance, we illustrate these approaches with three meta-analytic examples and present a step-by-step tutorial with open analytic code.

# Representing Primary Study Quality in Meta-Analyses

## Conceptualization of PSQ

Primary study quality means different things in different contexts. Although several definitions of the concept exist, these definitions, by and large, consider PSQ to be a characteristic of the extent to which a primary study's goals, design, and implementation match.[5] In this sense, PSQ addresses the validity of the inferences drawn from the study's results.[11] Moher emphasized that high-quality studies provide reproducible information and minimize the potential sources of bias in the study design, implementation, and the data analysis.[15,16] This emphasis links PSQ to the broader concept of "risk of bias".[4,17] Given that the potential sources of bias may include several study characteristics related to, for instance, the sampling (e.g., randomized vs. convenience sample), the outcome measures (e.g., reliability, evidence supporting a validity argument), the study design (e.g., randomized controlled trial vs. other designs), or the reporting of the results (e.g., reproducible reporting of statistical tests with analytic code, availability of the data), PSQ is a multidimensional concept and should be represented as such.[3] Several authors have criticized the mismatch between multidimensional definitions of PSQ and its representation as a single, unidimensional score.[5,12,18] Given the diversity of quality definitions, a plethora of assessment tools, scales, and checklists exist in the extant literature. These assessments have been adapted to several study designs (e.g., randomized controlled trials, cross-sectional studies) and research domains (e.g., healthcare, ecology, medicine),[2,19-23] and testify to the context-specificity of PSQ.[5]

## Key Decisions in Selecting or Creating PSQ Indicators and Scores

A common means of representing PSQ in a meta-analysis is to select or create quality indicators or scores based on several decision to best capture the meta-analyst's definition and operationalization of the concept. In this context, we draw from the psychometric terminology of measurement models[24] and refer "quality indicators" to the observed (manifest) variables describing PSQ at the levels of effect sizes, samples, studies, countries, time points etc.[25] These indicators are typically generated during the coding of the meta-analytic data. Furthermore, we refer "quality scores" to PSQ scores that have been aggregated from multiple quality indicators, for instance, as sum scores, average scores, or model-based factor scores.

As noted earlier, the ways in which meta-analysts have represented PSQ are manifold. In their systematic review of 225 meta-analyses that were published in the *Psychological Bulletin*, Wedderhoff and Bosnjak[7] identified three approaches meta-analysts have mainly taken to represent PSQ: (a) Selecting or creating a categorical indicator of PSQ based on one or more characteristics of the study design, methodology, or reporting; (b) Creating aggregated PSQ scores (e.g., model-based scale scores, a priori sum scores, composite scores) from multiple quality indicators; and (c) Using multiple quality indicators separately. To create such representations of PSQ in a meta-analysis, researchers have to make several analytic decisions. In our view, these key decisions include the (a) a priori vs. meta-analytic selection or creation of PSQ indicators or scores; (b) use of single vs. multiple quality indicators; (c) aggregation of multiple indicators into PSQ scores vs. keeping them separate; and (d) model-based vs. model-free PSQ score creation. We summarize these decisions in Figure 1 and explain them in greater detail in the following sections.

## A Priori vs. Meta-Analytic Selection of Creation of PSQ Indicators or Scores

Meta-analysts need to decide whether they select or create PSQ indicators or scores prior to or as part of the meta-analytic modeling. In the a priori methods, meta-analysts apply some conceptual and/or statistical criteria or procedures to select or create PSQ indicators or scores without any information about their possible relations to the effect sizes.[4] For instance, given the context of the meta-analysis, a meta-analyst may select study design characteristics, such as the design as a randomized-controlled trial or the random sampling of study participants, as quality indicators. Another meta-analysts may aggregate several PSQ indicators from a checklist into an aggregated score. These a priori methods are especially useful if meta-analysts wish to explore the nature of the relation between PSQ and the effect sizes.[26] However, these methods have at least two caveats: (a) They are based on the assumption that the a priori PSQ indicators or scores represent PSQ—an assumption that remains largely untested and that calls for crafting a validity argument;[11] and (b) a priori selected or created PSQ indicators or scores may not be predictive of the effect sizes[19]—in fact, they may not capture relevant PSQ aspects that are related to the effects[5]. In contrast, the meta-analytic methods identify the relations between PSQ and the effect sizes first and use this information in the selection or creation of PSQ indicators or scores. For instance, a meta-analyst may estimate the moderator effects of several study characteristics and select the indicators or scores that are actually related to the effect sizes as PSQ indicators. This method is especially useful if meta-analysts wish to create parsimonious meta-analytic models that incorporate PSQ and to maximize the heterogeneity explained by PSQ. Nevertheless, this meta-analytic method may result in different PSQ representations for different meta-analyses, thus challenging the comparability of the concept. Moreover, this method requires a sufficiently large, meta-analytic sample to estimate multiple moderator effects, for instance, via meta-regression models.[27]

### Single vs. Multiple Quality Indicators

Another key decision refers to the number of PSQ indicators. As noted earlier, some meta-analysts prefer to represent PSQ by a single indicator that describes, for instance, the study design (e.g., randomized-controlled trial vs. other designs). On the one hand, single PSQ indicators will result in parsimonious meta-analytic models, especially when meta-analysts wish to control for PSQ while studying other moderator effects, with straightforward interpretations of the PSQ moderator effects. Moreover, in the case of a single categorical PSQ indicators, meta-analysts can directly study the moderator effects on heterogeneity beyond the effect sizes.[14] On the other hand, single PSQ indicators may be too simplistic and little informative, given that they focus on one single aspect of the concept. In contrast, utilizing multiple PSQ indicators, however selected and coded, may represent several, relevant aspects of PSQ and may thus be more informative.[1,5,12] Besides, meta-analysts can also select among multiple PSQ indicators, applying conceptual and/or statistical criteria. However, as noted earlier, the more PSQ indicators are incorporated in meta-analytic models, the larger the meta-analytic sample must be.

### Keeping Indicators Separate vs. Score Aggregation

If meta-analysts decide for multiple PSQ indicators, another key decision to make is whether these indicators are kept separate or should be aggregated into a PSQ score. Keeping quality indicators separate has the advantage that meta-analysts can isolate which specific aspects of PSQ are significantly related to the effect sizes and explain heterogeneity. At the same time, utilizing multiple, separate PSQ indicators as possible moderators in meta-analytic models requires a sufficiently large meta-analytic sample[27] and a small degree of multicollinearity among the indicators.[28] In contrast, score aggregation methods reduce multiple PSQ indicators into one or more PSQ score(s) and thus require smaller meta-analytic samples for studying the relation between PSQ and the effect sizes. Although the resultant scores simplify the representation of PSQ, creating them has not been without criticism. For instance, Valentine argued that score aggregation can result in too simplistic representations of PSQ, through which distinct aspects of the concept are merged.[5] For instance, indicators of validity, reliability, and study design would be merged and equally weighted to create a PSQ score.[7] This merging may neglect the conceptual and empirical dimensionality of PSQ assessments, as it often results in a single PSQ sum score.[4,18] Moreover, the statistical assumptions underlying such sum score aggregation are rarely tested—thus, PSQ sum scores may or may not be valid.[9,10] In our view, PSQ score aggregation is especially useful to reduce the dimensionality of a set of PSQ indicators, given that the conceptual and empirical dimensionality have been examined and the resultant scores are interpretable.

### Model-Based vs. Model-Free Score Aggregation

If meta-analysts decide for aggregating multiple PSQ indicators into PSQ scores, they also have to decide which scores are created and how. Specifically, PSQ scores can be created using the information and parameters from statistical models (i.e., model-based) or without such information (i.e., model-free). In case of the latter, the validity of the resultant PSQ scores may be questionable, especially if there is no evidence backing the aggregation of multiple PSQ indicators. For instance, meta-analysts may create PSQ scores by counting the number of fulfilled quality criteria (i.e., PSQ indicators) in a checklist.[7] Although this count score has a straightforward interpretation as the number of fulfilled criteria, it may still not be meaningful, especially if the quality criteria represent different aspects of PSQ or some might be negatively correlated. Hence, aggregating multiple PSQ indicators into scores requires, at the minimum, an inspection of the conceptual and empirical associations among the indicators, irrespective of the decision for model-free or model-based aggregation methods. Specifically, inspecting the correlation matrix of the PSQ indicators is a key step towards selecting indicators for score aggregation and establishing the meaning of the resultant scores.[24]

The model-based score aggregation requires selecting a measurement model that represents the above-mentioned conceptual and empirical associations among the indicators. For instance, if all quality indicators are positively and at least moderately correlated, a reflective measurement model may be specified to extract one or more latent quality variables that are hypothesized to cause (co-)variation in the quality indicators.[24] Typical examples of this type of measurement models are common factor-analytic models (e.g., EFA and CFA) and models of item response theory (e.g., 1PL, 2PL, GPCM). Utilizing these models, meta-analysis can establish the meaning of the reflective latent variable (e.g., what is common among the PSQ indicators) and examine the evidence for construct validity. Unlike causal-formative and composite models, reflective models are easier to identify, for instance, by fixing the factor loadings or factor variances rather than incorporating additional outcome variables.[29] Hence, these types or measurement models are useful for creating a priori quality scores.[10]

Causal-formative or composite measurement models[30] are useful if the set of quality indicators contains both positive and negative correlations among the indicators.[24,31] In these models, the quality indicators are the hypothesized causes of the latent or composite quality variables[24,31,32]. However, these resultant variables do not have a meaning per se—the meta-analyst needs to ascribe a meaning to them.[33,34] Notably, causal-formative or composite measurement models are not identified, unless constraints on the factor loadings or weights are imposed or outcome variables are introduced that are explained by PSQ.[35] Hence, these models are useful when creating quality scores as part of the meta-analytic modeling.

## Incorporating Primary Study Quality in Meta-Analytic Models

Once meta-analysts have assessed, coded, and potentially calculated the quality of the eligible primary studies, they may now wish to utilize this information in their meta-analysis.

The extant literature describes several strategies through which information about PSQ can be incorporated in a meta-analysis. These strategies include *"a priori" methods*, in which PSQ informs the selection or inclusion of primary studies prior to the meta-analytic modeling and *"post-hoc" methods*, through which the impact of PSQ on the meta-analytic results are examined, for instance, via moderator or sensitivity analyses.[4] Beyond describing PSQ, a priori methods typically include but are not limited to[4,7,36]: (a) Using PSQ as an inclusion criterion for primary studies and thus restricting the meta-analytic evidence base to studies of a certain quality—this approach is not recommended due its limiting effects on the generalizability of the meta-analytic findings[26] and it should only be considered if PSQ is likely to be "important in the context of the research question"[5(p138)]; and (b) Weighting primary studies by their quality[37]—this approach is also not recommended due to the bias quality weights introduce into meta-analytic models[38]. Both of these a-priori methods have the potential to not only introduce biases but also lend hand to questionable research practices. For instance, the introduction of PSQ as inclusion and exclusion criteria can help to purposefully exclude primary studies that do not fit the expectations meta-analysts have stipulated.[39] Thus, the use of a priori methods should be cautioned.[5]

Post-hoc methods include but are not limited to:[4,7,36,40] (a) Analyzing publication bias in relation to PSQ—this way, meta-analysts can examine the extent to which primary studies of different quality may be more or less prone to publication bias;[36] (b) Conducting sensitivity analyses—this approach sheds light on the possible changes in meta-analytic findings when primary studies of a certain quality are excluded; and (c) Performing moderator analyses to describe the direct relation between PSQ and the effect sizes—this way, meta-analyses can compare the effects between studies with high vs. low quality via subgroup analyses[6] or estimate the moderator effects of PSQ via mixed-effects meta-regression models.[5,16]

Johnson and Valentine argued that post-hoc moderator analyses are preferable over the other methods, because they allow meta-analysts to explore the direct associations between PSQ and the effect sizes.[5,12] However, the current reports of these analyses in meta-analyses have several limitations: First, they are, by and large, limited to the linear moderator effects of continuous PSQ scores, although non-linear effects are conceptually plausible.[1] The lack of significant moderator effects of PSQ may be partly due to their non-linearity[1,5], and meta-analysts may actually expect non-linear relations (e.g., inverse quadratic relations with an "optimal" level of study quality for the largest effects and smaller effects for low- and high-quality studies). Second, moderator analyses have primarily focused on the direct moderator effects of PSQ on the effect sizes. However, recent extensions of random-effects models allow meta-analysts to explore these effects on heterogeneity estimates (e.g., mixed-effects models with heteroskedasticity for categorical PSQ scores, location-scale models for continuous PSQ scores).[13,14] Utilizing these model extensions provides information about the dependency of heterogeneity on PSQ. Third, the PSQ moderator effects are often reported independent of any other moderator. Nevertheless, using PSQ along with other moderators in, for instance, meta-regression models sheds light on the extent to which a moderator is related to the effect sizes after controlling for PSQ—that is, moderator compensation effects. Moreover, Johnson and colleagues[40] argued that meta-analysts should take an interactive approach to exploring the PSQ-effect size relation and estimate interactions between PSQ and other study characteristics—in this way, the sensitivity of moderator effects to PSQ can be explored.[12] Overall, the extensions of standard meta-analytic models allow meta-analysts to examine the possible impact of PSQ on meta-analytic findings in greater detail.

## Step-by-Step Tutorial

To facilitate representing PSQ and examining its moderator effects in meta-analyses, we propose taking five steps. Table 1 details these steps and outlines the analytic decisions within.

### Step 1: Study Quality Definition

First, we suggest defining PSQ in the context of the specific meta-analysis, given its research focus, selection criteria, and the characteristics of the meta-analytic data. Specifically, this step clarifies the aspects or dimensions of PSQ, such as the quality of the study design, sampling, measurement, and reporting,[4,41,42] and anchors PSQ in the frameworks of risk of bias and validity.[3,11] Another key element in the definition of PSQ is the specification of the conceptual level(s) at which the aspects or dimensions operate. For instance, meta-analysts may define PSQ at the levels of samples (e.g., representativeness of the sample), measurements (e.g., reliability coefficients), studies (e.g., study design features), or even countries (e.g., sampling quality and frame). Specifying these levels aids the interpretation of possible PSQ effects in subsequent meta-analytic models with multiple levels of analysis.[28]

### Step 2: Study Quality Operationalization

Second, we suggest operationalizing the definition of PSQ in a meta-analysis, that is, selecting observed quality indicators. This selection should not only be guided by the theoretical considerations behind the PSQ definition but also by empirical evidence. Specifically, we recommend inspecting the correlation matrix of quality indicators to identify whether they may indicate shared or unique quality constructs.[5] This information can then be used to specify, estimate, and evaluate a suitable measurement model of PSQ (i.e., reflective, causal-formative, or composite measurement model) to create quality scores.

**Step 3: Study Quality Score Creation**

Third, given the quality definition and operationalization, we recommend creating one or multiple quality scores as representatives of PSQ, either a priori or as part of the meta-analytic modeling. These scores can be single or multiple, categorical or continuous, aggregated or separate variables. As noted in the previous step, if meta-analysts decide to create a model-based quality score from multiple indicators a priori, then the measurement model they have decided on is now specified, estimated, and evaluated. For instance, meta-analysts may describe PSQ using a Rasch (one-parameter logistic) model or some categorical CFA model with a set of binary indicators from a well-established quality checklist. Typical a priori modeling approaches include but are not limited to models of item response theory, factor analysis, or principal component analysis. If this measurement model holds and represents the data well, then a sum or factor score can be extracted from it as a study quality score[9,43]. In essence, the third step we propose is to create the actual quality scores which can later be submitted to moderator or moderator sensitivity analyses.

**Step 4: Moderator Analyses**

Fourth, we recommend conducting moderator analyses to examine the extent to which study quality directly relates to the effect sizes[5,12]. These analyses yield information not only about the direct moderator effects of PSQ scores but also the extent to which heterogeneity can be explained at the respective level(s) of analysis identified in step 1[28]. Given the recent extensions of meta-analytic models to location-scale models or mixed-effects meta-regression models with subgroup-specific (residual) heterogeneities, meta-analysts can also explore possible moderator effects on the heterogeneity[13,14]. Moreover, moderator effects can be linear or curvilinear (e.g., quadratic with an inverse U-shape indicating smaller effects for low- and high-quality studies and a larger effect for some moderate quality score), and meta-analysts may explore the nature of the relations of PSQ with the effect sizes and their heterogeneity.

### Step 5: Moderator Sensitivity Analyses

Fifth, we recommend conducting moderator sensitivity analyses to examine the extent to which PSQ compensates or interacts directly with the effects of other moderators. Utilizing mixed-effects meta-regression models, meta-analysts can obtain information about the sensitivity of moderator effects to study quality.[12,44] For instance, the effect of a moderator may decrease or disappear after controlling for PSQ, and PSQ explains heterogeneity above and beyond the moderator. The effect of PSQ may be referred to as a "compensatory effect"[45]. Moreover, PSQ may directly interact with a moderator and, thus, the size of the moderator effect depends on the quality score[46]. Depending on the nature of the PSQ score and the moderator, centering may be required to circumvent multicollinearity issues in models containing PSQ, the moderator, and their interaction.[44]

## Illustrative Examples

To illustrate these steps and the analytic decisions within, we present three examples that showcase typical situations in which primary study quality is represented by (1) a priori selected, binary quality indicators; (2) a priori created, model-based quality scores; and (3) meta-analytically derived, model-based quality scores. We further highlight key analytic issues, such as meta-analytic model selection with binary moderators, criteria for selecting quality indicators that make up quality scores, factor structures of quality assessments, non-linear effects of study quality, creating composite quality scores from categorical and continuous indicators, and handling missing data in quality indicators. The illustrative examples are supplemented by open-access data, R code, and a detailed description of the results. Please find the respective material in the Open Science Framework at https://doi.org/10.17605/OSF.IO/NGVCZ.

## Illustrative Example 1: A Priori Selection of a Single Binary Quality Indicator

### Purpose and Context of the Example

The purpose of the first example is to illustrate how meta-analysts can use a binary quality indicator which is derived a priori, for instance, from categorizing a continuous quality score (e.g., low vs. high quality), summarizing multiple study characteristics (e.g., random assignment of participants to intervention groups in a pretest-posttest design vs. non-random assignment in a posttest-only design), or a single quality characteristic (e.g., random vs. convenience sampling). Specifically, going beyond standard mixed-effects models, we show a range of meta-analytic models that describe possible differences between low- and high-quality studies in the weighted average effect size *and* the respective heterogeneity[47].

The meta-analytic data set we use in this example was published by Scherer et al. (2019)[48] and contains 539 effect sizes from 105 primary studies and 31549 participants. The main goal of this meta-analysis was to examine the transfer effects of computer programming on cognitive skills. The authors selected (quasi-)experimental studies with a posttest-only or pretest-posttest experimental-control group design and included performance-based measures of cognitive skills.

### Steps 1-3: Study Quality Definition, Operationalization, and Score Creation

Given the focus of the meta-analysis on the possible transfer effects of computer programming interventions, characteristics of the study design seem a natural choice for representing study quality—hence, the level of primary studies is the conceptual level at which quality operates. In this illustrative example, we selected two variables—that is, the status of randomization (random vs. non-random) and the pretest-posttest experimental-control group (PPC) vs. posttest-only design—and combined them into a binary study quality score (labelled `binary.quality`). This score was coded as 1 for primary studies with a PPC design and a random assignment of participants to the intervention groups or 0 otherwise. In this sense, high-quality studies can be randomized-controlled trials or studies approximating them[49,50].

**Step 4: Moderator Analyses**

Examining possible moderator effects of the binary quality score needs to be based on a baseline model that represents the structure of the meta-analytic data. To find such a baseline model, we estimated a standard random-effects model and a three-level random-effects model that account for multiple effect sizes per study. We specified these models in the R package "metafor"[51] via the `rma.mv()` function. For the three-level model, we defined the random effects under the assumption of hierarchical effects (i.e., `random=list(~1|StudyID/ESID)`; see Table 2, Model 1)[28]. Comparing these two models suggested that the latter model was a better representation of the data than the former, as the likelihood-ratio test ($\chi^2[1] = 138.7$, $p < .001$) and the within- and between-study heterogeneity suggested ($\sigma_w^2 = 0.204$, $\sigma_b^2 = 0.281$). Adding information about a possible constant sampling correlation[52] between effect sizes within studies did not improve the model fit further. We therefore chose the three-level random-effects model with hierarchical effects as a baseline.

Next, we examined the direct moderator effects of study quality by extending the baseline model to a mixed-effects meta-regression model (see Model 2 in Table 2). In this extended model, the binary quality score did not show moderation effects and there was no evidence that high- and low-quality studies differed significantly in their weighted average effect sizes (high quality: $\bar{g} = 0.58$ vs. low quality: $\bar{g} = 0.46$; $F[1, 537] = 1.65$, $p = .20$). Comparing the baseline model to this model showed that PSQ explained only about 1.2% of between-study variation.

Notably, Model 2 assumed that high- and low-quality studies have the same amount of heterogeneity within and between studies after controlling for the differences in effect sizes (i.e., "residual heterogeneity"). However, this assumption may not be realistic, because either high- or low-quality studies may be more or less heterogeneous. Consequently, we further tested mixed-effects meta-regression models that allowed for different amounts of residual heterogeneity (see Models 3-5 in Table 2). Whilst several combinations of quality-specific parameters in the meta-analytic models are possible (i.e., same or quality-specific effect sizes, within-study variances, between-study variances; in total, 2×2×2 = 8 combinations), we focus on a selection of models in this tutorial (see Table 2). We argue that meta-analysts should decide for specific models on the basis of (a) theoretical considerations of where to reasonably expect differences in model parameters between high- and low-quality studies; and (b) comparisons of models with different assumptions on their parameters (e.g., via likelihood-ratio tests or an inspection of the information criteria)[53].

Ultimately, the model assuming quality-specific effect sizes and amounts of residual heterogeneity within and between studies (Model 5) outperformed the model assuming the same amounts of residual heterogeneity (Model 2), as the information criteria and the likelihood-ratio test suggested, $\chi^2(2) = 15.8$, $p < .001$. Further model comparisons (i.e., Model 5 vs. Models 3 and 4) indicated that Model 5 was also preferred over the other models (see Supplementary Material S1). Hence, although the effect sizes between high- and low-quality studies did not differ, the amounts of residual heterogeneity did. Overall, this example illustrates that PSQ may not only affect the overall effect sizes but also the heterogeneity.

### Step 5: Moderator Sensitivity Analyses

As another step, we conducted moderator compensation analyses to examine if PSQ compensated possible other moderator effects. We extended the mixed-effects meta-regression model assuming different amounts of residual heterogeneity within and between studies by the type of control group (i.e., `TreatedC`; coded as *1 = Control group was treated with another intervention targeted at improving the outcome*, *0 = Control group was untreated*)[48] as a moderator with or without controlling for study quality. The direct effect of the type of control group without controlling for study quality was $B = -0.48$ ($SE = 0.07$, $p < .001$). After controlling for study quality (via `mods=~factor(TreatedC)+factor(binary.quality)`), this effect remained, $B = -0.48$ ($SE = 0.07$, $p < .001$). The correlation between the study quality and the other moderator was small ($r = .03$, $p = .56$), so that multicollinearity was not an issue (VIF = 1.0). While only the moderator `TreatedC` explained about 24.6% of the within-study variation for low-quality studies and, respectively, 6.6% of the within-study variation of high-quality studies, accounted for differences in the weighted average effect sizes across study quality added only 0.2% and, respectively, 0.3% to this variance explanation. Hence, there was no evidence for a compensation effect of PSQ.

Finally, we tested whether study quality moderated the moderator effect of the type of control group by adding an interaction term between the two predictors (via `mods=~factor(TreatedC)*factor(binary.quality)`). There was no evidence for a statistically significant interaction ($B = 0.14$, $SE = 0.18$, $p < .001$), and comparing the models with and without the interaction term showed that adding the interaction did not improve the model fit ($\chi^2[1] = 2.2$, $p = .14$). Hence, high- and low-quality studies did not differ in the moderator effect of the type of control group—the moderator effect was not sensitive to PSQ.

## Illustrative Example 2: A Priori Model-Based Creation of a Study Quality Score

### Purpose and Context of the Example

The purpose of the second example is to illustrate how meta-analysts can create a study quality score a priori from a set of indicators. As noted earlier, quality assessments and checklists may provide such indicators, and researchers typically code a set of quality criteria as categorical or binary variables (e.g., the specific criterion is fulfilled or not).[12] On the basis of these variables, an overall quality score is created—oftentimes as the average rating across criteria, a sum or count score.[7] In this example, we show how meta-analysts can derive model-based quality scores from a set of positively correlated, binary indicators. The procedure we propose includes the checking of the indicator correlation matrix, the testing of reflective measurement models to create an overall quality score, and the checking for possible non-linear relations between primary study quality and the effect sizes.

The meta-analytic data set in this example was published by Siddiq and Scherer[54] and contains 53 effect sizes from 21 primary studies. These effect sizes represent the gender differences in students' digital skills which were measured by performance-based assessments and extracted from cross-sectional studies of 137,895 secondary-school students in 30 countries. Notably, this meta-analysis included aggregated summary data and individual-participant data.[55]

### Steps 1-3: Study Quality Definition, Operationalization, and Score Creation

In this illustrative example, we defined PSQ as a multidimensional concept describing the quality of the sampling, data, and measurement. Hence, primary study quality mainly operated at the levels of studies (e.g., sampling design) and effect sizes (e.g., measurement properties). The following a priori coded quality indicators were available in the meta-analytic data set (with their respective variable names and coding):

- Task interactivity (Interactivity; *1 = Interactive tasks*, *0 = Static tasks*);
- Assessment of applied skills (AppliedSkills; *1 = Tasks demanded applied skills*, *0 = Tasks demanded the recall of knowledge*);
- Random sampling (RandomSample; *1 = Random sampling*, *0 = Convenience sampling*);
- Test fairness evaluation (TestFair; *1 = Test fairness across gender groups assessed*, *0 = Not assessed*);
- Reliability reporting (RelReport; *1 = Scale reliability reported*, *0 = Not reported*);
- Availability of individual-participant data (IPD; *1 = IPD available*, *0 = Not available*).

The tetrachoric correlations among these quality indicators were positive and ranged between $r$ = 0.13 and $r$ = 0.89 (see Figure 2a). All indicators pointed into the same direction and shared some variation. Given the range of correlations, the set of indicators was likely multidimensional rather than unidimensional, and we examined the underlying factor structure via factor analysis. Depending on a priori hypotheses on the factor structure, meta-analysts may choose either exploratory or confirmatory factor analysis.

Using the tetrachoric correlation matrix, we found evidence for multidimensionality in the set of quality indicators. For instance, the Kaiser-Guttman criterion resulted in two eigenvalues above 1, the parallel analysis supported a two-factor solution, and the very simple structure analyses showed that an optimal fit to a simple structure was achieved already by two factors (see Supplementary Material S2). We therefore conducted an EFA with maximum-likelihood estimation as factor extraction and oblimin rotation as an oblique rotation, assuming two correlated factors. Using the `fa()` function in the R package “psych”[56] to estimate the EFA, we obtained a two-factor solution with one factor representing study design quality (indicated by TestFair, RandomSample, IPD, and Interactivity) and one factor representing measurement quality (indicated by RelReport and AppliedSkills). A subsequent CFA with WLSMV estimation and two correlated latent variables fitted the data well, $\chi^2(8) = 3.9$, $p = .86$. In sum, the factor analyses suggested that the set of quality indicators represents two study-quality scales, one indicating study design quality and one indicating measurement quality (see Figure 3).

As a next step, we explored the psychometric properties of the two study-quality scales using item response theory (IRT) models. IRT models are probabilistic models describing the non-linear (logistic) link between the probability of choosing a specific response category (e.g., correct vs. incorrect, Likert scales) and an underlying construct, the latent trait. This link can further be described by additional item parameters, such as item difficulties or discriminations.[43] We used IRT to describe PSQ due its efficient modeling of latent traits based on categorical indicators, the low demands on sample size, and the invariance of item parameters and latent traits.[57,58] For a detailed description of these models and their underlying assumptions, we refer readers to de Ayala.[43] Notably, we assumed that all quality indicators share common variance which can be captured by a reflective latent variable. We estimated the IRT models using the `tamaan()` function in the R package “TAM”[59] and the `mirt()` function in the R package “mirt”.[60]

Drawing from the evidence on two correlated quality scales (see Figure 3), we estimated both unidimensional and two-dimensional IRT models with one or two item parameters (1PL and 2PL; see Supplementary Material S2). While the former assumed equal item discriminations (factor loadings) across quality indicators, the latter estimated them freely. Table 3 shows the information criteria and goodness-of-fit indices of the respective models. For both the 1PL and the 2PL models, the multidimensional models were preferred over the unidimensional models. Moreover, among the four IRT models, the two-dimensional 2PL model was preferred—hence, we chose this model to represent the two study-quality scales. The two latent variables were moderately correlated ($\rho$ = .71) and, given the few quality indicators, exhibited substantial marginal reliabilities (study design quality: $\hat{r}_{xx}$ = .70, measurement quality: $\hat{r}_{xx}$ = .55). The corresponding test information curve is shown in Figure 2b and indicated that the quality assessment provided the most information and best precision when the values of both factors were between -2.0 and 0.0. Supplementary Material S2 contains further details of this model, such as the parameter estimates of difficulty and discrimination, local independence indices, item and person fit information, and the item characteristic curves of the quality indicators. Finally, we extracted the expected-a-posteriori (EAP) factor scores from the two-dimensional 2PL model via the `IRT.factor.scores()` function and used them as study quality scores.

### Step 4: Moderator Analyses

Similar to the first illustrative example, we selected a meta-analytic baseline model on the basis of model comparisons and estimated variance components (see Supplementary Material S2). We accepted a three-level random-effects model with samples nested in primary studies as the baseline model for further moderator analyses. Examining the possible linear moderator effects of the two study-quality scores separately via mixed-effects meta-regression models, we did not find evidence for significant moderation by study design quality ($B$ = 0.06, *SE* = 0.05, *p* = .19; see Figure 4) or measurement quality ($B$ = 0.06, *SE* = 0.04, *p* = .12) in these models. However, when using robust standard errors generated with the R package “clubSandwich”[61], the linear moderator effect of measurement quality was significant ($B$ = 0.06, *SE* = 0.02, *p* = .03) and suggested that primary studies with high measurement quality tended to show larger effect sizes. The joint moderator effects of the two quality scores and their interaction were insignificant (see Supplementary Material S2).

As a next step, we explored possible non-linear effects by including up to cubic polynomial moderator effects of the quality scores (e.g., `mods=~poly(zEAP2PLFS1, degree=3)` with `zEAP2PLFS1` representing the *z*-standardized EAP factor score from the 2PL model representing measurement quality). While there was no evidence for such effects of measurement quality, there was a tendency toward a cubic relation between the effect sizes and study design quality (see Figure 4). Please refer to the Supplementary Material S2 for the detailed estimates. Overall, about 30.4% of the between-study variation could be explained by the non-linear effects of study design quality.

### Step 5: Moderator Sensitivity Analyses

We illustrate the moderator compensation and interaction effects using the Power Distance Index (PDI) in the data set as an example moderator. PDI showed a negative moderator effect (*B* = -0.04, *SE* = 0.02, *p* = .05; *F*[1, 51] = 4.00, *p* = .05), explained about 9.6% of the within-study variation (0% of the between-study variation), and indicated that samples in countries with larger power distance tended to show smaller effect sizes. After controlling for study design quality, this effect remained (*B* = -0.04, *SE* = 0.02, *p* = .03; *F*[2, 50] = 3.26, *p* = .05). Similarly, the effect of PDI remained after controlling for measurement quality (*B* = -0.04, *SE* = 0.02, *p* = .03; *F*[2, 50] = 3.65, *p* = .03). PDI and study design quality explained about 22.3% of the between-study variation and about 4.9% of the within-study variation—similarly, PDI and measurement quality explained about 19.6% and 5.3%, respectively. Controlling for study design and measurement quality jointly also did not affect the moderation effect by PDI (*B*

= -0.04, *SE* = 0.02, *p* = .03; *F*[3, 49] = 2.43, *p* = .08). Overall, above and beyond PDI, the study quality scores explained variation in the effect sizes—hence, PSQ showed a compensation effect.

Finally, we examined the interaction between study quality and the moderator PDI (e.g., `mods=~zEAP2PLFS1*scale(PDI)`). While there was no interaction effect with study design quality (*B* = 0.04, *SE* = 0.04, *p* = .23), there was a significant interaction with measurement quality (*B* = 0.10, *SE* = 0.05, *p* = .06). In this sense, the moderator effect of PDI was sensitive to measurement quality and tended to increase with higher quality scores.

## Illustrative Example 3: Meta-Analytic Model-Based Creation of a Study Quality Score

### Purpose and Context of the Example

The purpose of the third example is to illustrate how meta-analysts can represent primary study quality by a composite score that is derived from a set of mixed-format indicators, either negatively or positively correlated (e.g., reliability coefficients, status of openly available data, study design characteristics). Moreover, this composite score is informed by a mixed-effects meta-regression model that contains the quality indicators as simultaneous moderators. Given that quality indicators were missing for some studies, we conducted sensitivity analyses, comparing the results of the incomplete and multiply imputed data. We used the same meta-analytic data set as in the first illustrative example.

### Steps 1-3: Study Quality Definition, Operationalization, and Score Creation

Given that the meta-analysis was concerned with the effectiveness of interventions, we defined primary study quality as a concept that is comprised of information about the study design and sampling, the publication status, and the measurement quality of the outcome variable. Study quality contains this information at the level of effect sizes and primary studies. To operationalize study quality, we chose the following quality indicators:

- Publication status (PubStatus; *1 = Published study, 0 = Grey literature*);
- Intervention design (PPCDesign; *1 = Pretest-posttest experimental-control group design, 0 = Posttest-only experimental-control group design*);
- Group assignment (Random; *1 = Randomization of the group assignment, 0 = Non-random assignment*);
- Treatment of the control group(s) (TreatedC; *1 = Treated control group, 0 = Untreated control group*);
- Matching of the groups (Matched; *1 = Matching of the groups was performed, 0 = No matching was performed*);
- Standardized performance assessment of the outcome (PerfSTA; *1 = Performance-based, standardized assessments, 0 = Performance-based, self-developed, or non-standardized assessments*); and
- Reported reliability coefficient of the assessment (Reliability).

Some of the primary studies did not provide any reports on the matching of groups (43.2% missing) or the reliability of the outcome (3.0% missing). To handle these missing data, we performed multiple imputation and generated 20 complete data sets. Specifically, we used the multiple imputation with chained equations procedure implemented in the R package "mice"[62] and imputed the binary variable "Matched" via logistic regression (`logreg`) and the continuous variable "Reliability" via predictive mean matching (`pmm`). We performed all subsequent analyses for each of the 20 imputed data sets and combined the resultant parameters following Rubin's combination rules.[63] For more details on multiple imputation in meta-analysis, please refer to Viechtbauer[64] and Lee and Beretvas[65]. To show the possible effects this missing data treatment may have, we present both the incomplete and imputed data analyses in the

Supplementary Material S3. In the following sections, we only present the results obtained from the multiply imputed data.

To examine the correlations among the binary and continuous variables in the meta-analytic data set, we estimated the correlations using the `mixedCor()` function in the R package "psych".[56] The resultant correlation matrix is shown in Figure 5a. Notably, the quality indicators exhibited positive, negative, and close-to-zero correlations and ranged between *r* = -.30 and *r* = .43, despite that they were coded with the same direction (i.e., consistently across all indicators, high indicator scores suggested better study quality). Moreover, the correlations between the effect sizes and the quality indicators were similarly diverse, with a range between *r* = -.18 and *r* = .24. In a situation with mixed correlations, meta-analysts who wish to utilize all quality indicators have at least two analytic options: (a) using all quality indicators but keeping them separate in further analyses; or (b) creating a composite score. While we focus on the latter in this illustrative example, we also showcase the former to allow meta-analysts to compare the meta-analytic results. Hence, our approach to representing PSQ is to create a composite quality score.

Given that composite measurement models are not identified without one or more additional outcome variables,[24,30] we created the composite score in two steps (see Figure 6): First, we estimated a mixed-effects meta-regression model with all quality indicators as moderators of the effect sizes, yet without an intercept. This step generated a set of weights $w_i$, one for each quality indicator $X_{ij}$. Second, we created the composite quality score $C_j$ for each study $j$ as the weighted sum of all quality indicators, $C_j = \sum_{i=1}^{7} w_i X_{ij}$.

As noted in illustrative example 1, a three-level random-effects model formed the meta-analytic baseline model. To generate the weights, we extended this model by adding the quality indicators as moderators (`mods=~0+PubStatus+…`), extracted the resultant regression coefficients, and created the composite quality score (see Supplementary Material S3). Unlike in the factor scores representing study quality illustrative example 2, this score cannot be interpreted as a typical sum score with high values indicating better quality. Instead, the composite quality score captures the possible moderator effects of quality indicators, into whichever direction they may point, and aggregates them in one score.

### Step 4: Moderator Analyses

**Separate quality indicators.** For both the incomplete and the multiply imputed data sets, we found that the treatment of control groups and the type of performance-based assessment were significantly related to the treatment effects after controlling for all other quality indicators (see Supplementary Material S3). Table 4 shows the respective regression parameters for the multiply imputed data. Specifically, primary studies with treated control groups had smaller intervention effects, and primary studies with standardized performance assessments had smaller effects than studies with self-developed and non-standardized assessments. The variance inflation factors and the correlations among the quality indicators did not point to a possible multicollinearity issue in the meta-regression models (VIFs < 2.0; see Figure 5a and Supplementary Material S3). Overall, about 16.1% of the between-study variation could be explained by the set of quality indicators.

**Composite quality score.** The composite quality score exhibited a linear relation to the effect size, with some deviations (see Figure 5b). The respective regression coefficient was positive, and, similar to the variance explained by the model keeping the quality indicators separate, the quality score explained about 18.9% of the between-study variation (see Table 4).

### Step 5: Moderator Sensitivity Analyses

**Separate quality indicators.** Studies that focused on creativity as an outcome variable ("Creativity"; coded as *1 = Creativity assessment, 0 = Assessment of other cognitive skills*) tended to show larger intervention effects than those focusing on other cognitive skills ($B$ = 0.38, $SE$ = 0.12, $p$ < .01). This moderator effect remained after controlling for the quality indicators ($B$ = 0.41, $SE$ = 0.13, $p$ < .01; see Supplementary Material S3). Adding study quality to the mixed-effects meta-regression model explained about 18.8 % of the within-study variation, while the moderator Creativity explained about 3.5% without the quality indicators and the quality indicators without Creativity explained about 16.1%. In this sense, controlling for study quality compensated the moderator effect of Creativity only marginally. Moreover, none of the interaction terms between Creativity and the quality indicators were statistically significantly different from zero, so that the moderator effect was not sensitive to the single quality indicators (see Supplementary Material S3).

**Composite quality score.** Similar to keeping the quality indicators separate, the moderator effect of Creativity remained after controlling for the composite quality score ($B$ = 0.35, $SE$ = 0.12, $p$ < .01). Creativity and study quality explained about 22.1% of the within-study variation, while study quality alone explained 18.9%. Hence, the Creativity effect was not affected by controlling for study quality. At the same time, the moderator effect of Creativity tended to be larger for studies with larger quality scores (interaction effect: $B$ = 0.21, $SE$ = 0.12, $p$ = .07; see Supplementary Material S3). In other words, the Creativity effect was sensitive to study quality.

## General Discussion

### Implications for Meta-Analytic Practice

Utilizing information about the quality of primary studies in meta-analyses supports meta-analysts in their crafting of a validity argument. Quality information further provides contextual information through which weighted average effect sizes and heterogeneity could be interpreted and explored. Best-practice guidelines consequently consider the coding, selection, and creation of PSQ indicators and scores to be key steps in any meta-analysis.[66,67] Our tutorial and the illustrative examples were aimed at supporting best meta-analytic practices by providing hands-on guidance in these steps.

First, selecting or creating quality indicators and scores a priori or as part of the meta-analytic modeling is a key decision meta-analysts have to make. In our view, this decision largely depends on the purpose of quantifying PSQ in a meta-analysis. If meta-analysts wish to map the quality of primary studies in their meta-analysis onto existing research, an a priori generation of quality scores using well-established checklists and assessments could ensure comparability to other meta-analyses.[4,19,68,69] Moreover, an a priori generation of quality scores allows meta-analysts to explore the relations between PSQ and the effect sizes irrespective of prior information about the direction, strength, and statistical significance of this relation. In contrast, quality scores that are generated or selected as part of the meta-analytic modeling contain information about these relations and are thus useful when meta-analysts wish to maximize the amount of heterogeneity explained by PSQ.

Second, using a single or multiple quality indicators or scores is another key decision meta-analysts have to make when representing PSQ. Selecting or creating these indicators operationalizes PSQ and should thus be based on substantive theories and the context in which the meta-analysis is conducted.[70] For instance, quality indicators of randomized-controlled trials may differ from those describing cross-sectional surveys, and accounting for this context-dependency is key to measuring PSQ.[4] Despite the analytic advantages (e.g., straightforward interpretation and modeling of moderator effects), meta-analysts should be aware of the limitations of selecting or creating a single quality indicator, such as a limited representation of PSQ and a possible loss of quality information due to the aggregation of multiple indicators in a single indicator or score.[7] Following Johnson's recommendations,[12] we argue for a multidimensional representation of PSQ, with multiple indicators capturing multiple quality aspects and dimensions. In this way, nuanced information about which quality aspects may or may not explain heterogeneity is obtained.

Third, if meta-analysts wish to aggregate multiple quality indicators into a single quality score or few quality scores, they need to decide on the approach through which this aggregation is achieved—model-based or model-free. For instance, it has been a standard practice to use established quality checklists and create sum scores to represent PSQ.[4,7] While hardly any statistical modeling is involved in this practice, creating a sum score is based on assumptions about a statistical model underlying the checklist.[9] Despite the availability of quality checklists and assessments, we still consider it important to check these underlying assumptions and to obtain reliability and validity evidence. This includes the checking of psychometric properties of the quality indicators, along with the selection and testing of appropriate measurement models.[10,24] Using the resultant information, meta-analysts could disclose the properties of the quality assessment in their meta-analysis and, ultimately, craft a validity argument for the generated quality scores.

Fourth, in our view, examining possible moderator effects of study quality should go beyond testing linear relations between PSQ scores and the effect sizes that might explain heterogeneity. We suggest considering possible non-linear moderator effects of PSQ, moderator effects on effect sizes *and* their heterogeneity, and the sensitivity of other moderator effects to PSQ.[12] In this way, meta-analysts can shed further light on the impact of study quality on the meta-analytic findings beyond linear moderation.

## Limitations and Future Research Directions

The methodological approaches we have presented have some limitations. First, the direct integration between reflective measurement models describing PSQ at the levels of effect sizes and studies and meta-analytic models is still to be implemented. Currently, the a priori model-based creation of PSQ scores based on multiple quality indicators and the use of these scores in moderator (sensitivity) analyses are disconnected and require two analytic steps. The existing approaches of meta-analytic and multilevel structural equation modeling[53,55] may provide frameworks for developing a one-step approach. Second, the model-based approaches to creating PSQ scores are based on substantive and statistical assumptions, limitations, and interpretations meta-analysts need to be aware of. For instance, composite quality scores that are based on negatively, positively, and zero correlated indicators do not have a clear-cut interpretation, such as “the higher the score, the better the quality”. This requires that meta-analysts clearly and transparently communicate their decisions and the reasoning behind them. We therefore encourage the development of open science practices and standards for evaluating PSQ in meta-analyses.[4] Third, although our step-by-step guidance was designed for meta-analyses, the proposed steps 1-3 apply to systematic reviews without a quantitative synthesis of effect sizes. Moreover, the approaches to create or select PSQ indicators or scores a priori, conceptually or model-based, are accessible as well. Researchers can then utilize this information to critically appraise the primary study quality.[21] Fourth, the model-based approaches require extra time in the long meta-analytic process and some methodological expertise beyond standard meta-analysis. In our view, spending this time and acquiring this expertise is key to representing and utilizing PSQ in meta-analysis by indicators from existing quality checklists and assessments.

## Conclusions

Our review, discussion, and illustration of representing and utilizing PSQ in meta-analyses showed that a variety of approaches to select and create quality indicators and scores exist. These approaches vary in the level of information they provide, the assumptions they make, and the usefulness in meta-analytic models. While a key strength of binary quality scores lies in their ease of interpretation and the possible range of differences between low- and high-quality studies in the meta-analytic model parameters that can be explored, categorizing study quality is often associated with a loss of information about the variation in quality. Model-based approaches can address this issue and combine several quality indicators into a continuous quality score—however, the interpretation of such a score largely depends on the correlational structure of the underlying indicators and the measurement model it was generated from.

First and foremost, we conclude that meta-analysts must decide whether the quality indicators or scores are selected or created a priori or as part of the meta-analytic model. This decision is primarily driven by the purpose of including primary study quality in a meta-analysis, the substantive meaning of study quality, and the available quality indicators. Second, we conclude that PSQ should be clearly defined, including the level at which the concept operates. Third, we argue that simply averaging or summing up scores from a quality assessment or checklist without testing the prerequisites for such practices threatens the validity of the resultant quality scores. We encourage meta-analysts to consider the model-based creation of quality scores when multiple indicators are to be combined. Finally, we conclude that the analysis of moderation by study quality should go beyond direct moderator effects and include both moderator compensation and sensitivity analyses.

## Highlights

### What is already known

- Assessing and mapping primary study quality is a key step in meta-analyses to indicate the possible risk of bias.
- Primary study quality is often represented by single quality scores which are either based on categorical ratings of some quality criterion or the sum of multiple criteria.
- To examine the extent to which primary study quality influences the meta-analytic results, meta-analysts can explore the moderator effects of the respective quality scores.

### What is new

- Primary study quality scores can be selected or created either a priori or as part of the meta-analytic modeling.
- Creating a priori quality scores from multiple quality indicators (e.g., checklists and assessments with multiple criteria) requires selecting and evaluating suitable measurement models.
- Model-based meta-analytic quality scores allow meta-analysts to combine in a single composite score the information about the relations between multiple quality criteria and their moderator effects.
- Examining the impact of study quality includes estimating its moderator effects, the effects of other moderators after controlling for study quality, and the interactions between study quality and other moderators.

### Potential impact for *Research Synthesis Methods* readers

- This tutorial guides researchers in representing *and* incorporating primary study quality in their meta-analysis by describing, reviewing, and illustrating the respective analytic decisions and approaches.
- We provide step-by-step guidance, illustrative examples of openly available data sets, and hands-on R code.

## References

1. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nurs Res*. 1991;40(6):352-5.

2. Buccheri RK, Sharifi C. Critical Appraisal Tools and Reporting Guidelines for Evidence-Based Practice. *Worldviews on Evidence-Based Nursing*. 2017;14(6):463-472. doi:10.1111/wvn.12258

3. Feeley TH. Assessing Study Quality in Meta-Analysis. *Human Communication Research*. 2020;46(2-3):334-342. doi:10.1093/hcr/hqaa001

4. Hohn RE, Slaney KL, Tafreshi D. Primary Study Quality in Psychological Meta-Analyses: An Empirical Assessment of Recent Practice. *Frontiers in Psychology*. 2019;9doi:10.3389/fpsyg.2018.02667

5. Valentine JC. Incorporating judgments about study quality into research syntheses. In: Cooper H, Hedges L, Valentine J, eds. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation; 2019:129-140:chap 7.

6. Khan KS, Daya S, Jadad AR. The Importance of Quality of Primary Studies in Producing Unbiased Systematic Reviews. *Archives of Internal Medicine*. 1996;156(6):661-666. doi:10.1001/archinte.1996.00440060089011

7. Wedderhoff N, Bosnjak M. Erfassung der Primärstudienqualität in psychologischen Meta-Analysen. *Psychologische Rundschau*. 2020;71(2):119-126. doi:10.1026/0033-3042/a000484

8. Widaman KF, Revelle W. Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*. 2023/02/01 2023;55(2):788-806. doi:10.3758/s13428-022-01849-w

9. McNeish D, Wolf MG. Thinking twice about sum scores. *Behavior Research Methods*. 2020/12/01 2020;52(6):2287-2305. doi:10.3758/s13428-020-01398-0

10. Albanese E, Bütikofer L, Armijo-Olivo S, Ha C, Egger M. Construct validity of the Physiotherapy Evidence Database (PEDro) quality scale for randomized trials: Item response theory and factor analyses. *Research Synthesis Methods*. 2020;11(2):227-236. doi:10.1002/jrsm.1385

11. Valentine JC. Judging the quality of primary research. In: Cooper H, Hedges LV, Valentine JC, eds. *The handbook of research synthesis and meta-analysis*. 2nd ed. Russell Sage Foundation; 2009:129-146.

12. Johnson BT. Toward a more transparent, rigorous, and generative psychology. *Psychological Bulletin*. 2021;147(1):1-15. doi:10.1037/bul0000317

13. Viechtbauer W, López-López JA. Location-scale models for meta-analysis. *Research Synthesis Methods*. 2022/11/01 2022;13(6):697-715. doi:10.1002/jrsm.1562

14. Rubio-Aparicio M, López-López JA, Viechtbauer W, Marín-Martínez F, Botella J, Sánchez-Meca J. Testing Categorical Moderators in Mixed-Effects Meta-analysis in the Presence of Heteroscedasticity. *The Journal of Experimental Education*. 2020;88(2):288-310. doi:10.1080/00220973.2018.1561404

15. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials*. 1995;16(1):62-73. doi:10.1016/0197-2456(94)00031-W

16. Conn VS, Rantz MJ. Research methods: Managing primary study quality in meta-analyses. https://doi.org/10.1002/nur.10092. *Research in Nursing & Health*. 2003/08/01 2003;26(4):322-333. doi:https://doi.org/10.1002/nur.10092

17. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928. doi:10.1136/bmj.d5928

18. Valentine J, Cooper H. A Systematic and Transparent Approach for Assessing the Methodological Quality of Intervention Effectiveness Research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*. 2008;13(2):130-149. doi:10.1037/1082-989X.13.2.130

19. Brouwers MC, Johnston ME, Charette ML, Hanna SE, Jadad AR, Browman GP. Evaluating the role of quality assessment of primary studies in systematic reviews of cancer practice guidelines. *BMC Medical Research Methodology*. 2005;5(1):8. doi:10.1186/1471-2288-5-8

20. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools. *Journal of Evaluation in Clinical Practice*. 2019;25(1):44-52. doi:10.1111/jep.12889

21. Stanhope J, Weinstein P. Critical appraisal in ecology: What tools are available, and what is being used in systematic reviews? *Research Synthesis Methods*. 2023;14(3):342-356. doi:10.1002/jrsm.1609

22. Tran L, Tam DNH, Elshafay A, Dang T, Hirayama K, Huy NT. Quality assessment tools used in systematic reviews of in vitro studies: A systematic review. *BMC Medical Research Methodology*. 2021/05/08 2021;21(1):101. doi:10.1186/s12874-021-01295-w

23. Ma L-L, Wang Y-Y, Yang Z-H, Huang D, Weng H, Zeng X-T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*. 2020;7(1):7. doi:10.1186/s40779-020-00238-8

24. Kline RB. Assumptions in Structural Equation Modeling. In: Hoyle R, ed. *Handbook of Structural Equation Modeling*. 2nd ed. Guilford Press; 2023:128-144:chap 7.

25. Zeng X, Zhang Y, Kwong JSW, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. *Journal of Evidence-Based Medicine*. 2015;8(1):2-10. doi:10.1111/jebm.12141

26. Carroll C, Booth A. Quality assessment of qualitative evidence for systematic review and synthesis: Is it meaningful, and if so, how should it be performed? *Research Synthesis Methods*. 2015;6(2):149-154. doi:10.1002/jrsm.1128

27. Tipton E, Pustejovsky JE, Ahmadi H. Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*. 2019/06/01 2019;10(2):180-194. doi:10.1002/jrsm.1339

28. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. *Doing Meta-Analysis in R: A Hands-on Guide*. PROTECT Lab.

29. Kline RB. *Principles and practice of structural equation modeling, 4th ed*. Principles and practice of structural equation modeling, 4th ed. Guilford Press; 2016:xvii, 534-xvii, 534.

30. Henseler J. *Composite-Based Structural Equation Modeling: Analyzing Latent and Emergent Variables*. Guilford Press; 2021.

31. Iris B, Laura S, Dennis B. Should indicators be correlated? Formative indicators for healthcare quality measurement. *BMJ Open Quality*. 2022;11(2):e001791. doi:10.1136/bmjoq-2021-001791

32. Hempel S, Booth M, Miles J, et al. *Empirical Evidence of Associations Between Trial Quality and Effect Sizes. Methods Research Report*. 2011. *Methods Research Reports*. https://www.ncbi.nlm.nih.gov/books/NBK56925/

33. Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. *Psychological Review*. 2003;110(2):203-219. doi:10.1037/0033-295X.110.2.203

34. Edwards JR. The Fallacy of Formative Measurement. *Organizational Research Methods*. 2010;14(2):370-388. doi:10.1177/1094428110378369

35. Bollen KA, Bauldry S. Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*. 2011;16(3):265-284. doi:10.1037/a0024448

36. Sterne JAC, Hernán MA, McAleenan A, Reeves BC, Higgins JPT. Assessing risk of bias in a non-randomized study. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. The Cochrane Collaboration and John Wiley & Sons Ltd.; 2019:621-642:chap 25.

37. Rosenthal R. Quality-weighting of studies in meta-analytic research. *Psychotherapy Research*. 1991/07/01 1991;1(1):25-28. doi:10.1080/10503309112331334031

38. Ahn S, Becker BJ. Incorporating Quality Scores in Meta-Analysis. *Journal of Educational and Behavioral Statistics*. 2011;36(5):555-585. doi:10.3102/1076998610393968

39. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*. 2012;23(5):524-532. doi:10.1177/0956797611430953

40. Johnson BT, Low RE, MacDonald HV. Panning for the gold in health research: Incorporating studies' methodological quality in meta-analysis. *Psychology & Health*. 2015;30(1):135-152. doi:10.1080/08870446.2014.953533

41. Protogerou C, Hagger MS. A checklist to assess the quality of survey studies in psychology. *Methods in Psychology*. 2020;3:100031. doi:10.1016/j.metip.2020.100031

42. Brown SA. Measurement of quality of primary studies for meta-analysis. *Nursing Research*. 1991;40(6):352-355.

43. de Ayala RJ. *The Theory and Practice of Item Response Theory*. 2nd ed. The Guilford Press; 2022.

44. Knop ES, Pauly M, Friede T, Welz T. The consequences of neglected confounding and interactions in mixed-effects meta-regression: An illustrative example. https://doi.org/10.1002/jrsm.1643. *Research Synthesis Methods*. 2023/06/04 2023;n/a(n/a)doi:https://doi.org/10.1002/jrsm.1643

45. Scherer R. The Case for Good Discipline? Evidence on the Interplay Between Disciplinary Climate, Socioeconomic Status, and Science Achievement from PISA 2015. In: Frønes TS, Pettersen A, Radišić J, Buchholtz N, eds. *Equity, Equality and Diversity in the Nordic Model of Education*. Springer International Publishing; 2020:197-224.

46. Li X, Dusseldorp E, Meulman JJ. Meta-CART: A tool to identify interactions between moderators in meta-analysis. *British Journal of Mathematical and Statistical Psychology*. 2017;70(1):118-136. doi:10.1111/bmsp.12088

47. Rubio-Aparicio M, Sánchez-Meca J, López-López JA, Botella J, Marín-Martínez F. Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances. *British Journal of Mathematical and Statistical Psychology*. 2017;70(3):439-456. doi:10.1111/bmsp.12092

48. Scherer R, Siddiq F, Sánchez Viveros B. The cognitive benefits of learning computer programming: A meta-analysis of transfer effects. *Journal of Educational Psychology*. 2019;111(5):764-792. doi:10.1037/edu0000314

49. Sullivan GM. Getting Off the "Gold Standard": Randomized Controlled Trials and Education Research. *Journal of Graduate Medical Education*. 2011;3(3):285-289. doi:10.4300/JGME-D-11-00147.1

50. Torgerson CJ, Torgerson DJ. The Need for Randomised Controlled Trials in Educational Research. *British Journal of Educational Studies*. 2001;49(3):316-328. doi:10.1111/1467-8527.t01-1-00178

51. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*. 2010;36(3)

52. Pustejovsky JE, Tipton E. Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*. 2021;doi:10.1007/s11121-021-01246-3

53. Cheung MW-L. *Meta-Analysis: A Structural Equation Modeling Approach*. John Wiley & Sons Ltd.; 2014.

54. Siddiq F, Scherer R. Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational Research Review*. 2019;27:205-217. doi:10.1016/j.edurev.2019.03.007

55. Campos DG, Cheung MW-L, Scherer R. A Primer on Two-Stage Meta-Analysis with Individual Participant Data Obtained from Complex Sampling Surveys. *Psychological Methods*. 2023;doi:10.1037/met0000539

56. *psych: Procedures for Psychological, Psychometric, and Personality Research*. 2023. https://CRAN.R-project.org/package=psych

57. Finch H, French BF. A Comparison of Estimation Techniques for IRT Models With Small Samples. *Applied Measurement in Education*. 2019/04/03 2019;32(2):77-96. doi:10.1080/08957347.2019.1577243

58. Wu M, Tam HP, Jen T-H. *Educational Measurement for Applied Researchers: Theory into Practice*. Springer; 2016.

59. *TAM: Test Analysis Modules. R package version 4.1-4*. 2022. https://CRAN.R-project.org/package=TAM

60. Chalmers RP. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*. 2012;48(6):1 - 29. doi:10.18637/jss.v048.i06

61. *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.5.3*. 2021. https://CRAN.R-project.org/package=clubSandwich

62. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1 - 67. doi:10.18637/jss.v045.i03

63. Enders CK. *Applied missing data analysis*. 2nd ed. Applied missing data analysis. Guilford Press; 2022.

64. Viechtbauer W. Multiple Imputation with the mice and metafor Packages. Updated 03 August 2022. Accessed 19 April, 2023. https://www.metafor-project.org/doku.php/tips:multiple_imputation_with_mice_and_metafor

65. Lee J, Beretvas SN. Comparing methods for handling missing covariates in meta-regression. *Research Synthesis Methods*. 2023;14(1):117-136. doi:10.1002/jrsm.1585

66. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: The APA Publications and

Communications Board task force report. *American Psychologist*. 2018;73(1):3-25. doi:10.1037/amp0000191

67. Pigott TD, Polanin JR. Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*. 2020/02/01 2020;90(1):24-46. doi:10.3102/0034654319877153

68. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *Journal of Evaluation in Clinical Practice*. 2012;18(1):12-18. doi:10.1111/j.1365-2753.2010.01516.x

69. De Santis KK, Lorenz RC, Lakeberg M, Matthias K. The application of AMSTAR2 in 32 overviews of systematic reviews of interventions for mental and behavioural disorders: A cross-sectional study. *Research Synthesis Methods*. 2022;13(4):424-433. doi:10.1002/jrsm.1532

70. Schang L, Blotenberg I, Boywitt D. What makes a good quality indicator set? A systematic review of criteria. *International Journal for Quality in Health Care*. 2021;33(3):mzab107. doi:10.1093/intqhc/mzab107

71. Pavlov G, Maydeu-Olivares A, Shi D. Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models. *Educational and Psychological Measurement*. 2020;81(1):110-130. doi:10.1177/0013164420926231

72. Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. Sage Publications; 2012.

## Tables

**Table 1**

*Step-by-Step Guide to Representing and Incorporating Primary Study Quality in Meta-Analyses*

| Step | Description |
|---|---|
| **Step 1:** Study Quality Definition | ▪ Definition of the **aspects or dimensions of PSQ** (e.g., quality of the study design, sampling, measurement, or reporting)<br>▪ Definition of the **conceptual level(s)** at which the quality aspects or dimensions operate (e.g., level of the measures, samples, studies, or countries) |
| **Step 2:** Study Quality Operationalization | ▪ Conceptual and empirical **selection of quality indicators**, that is, coded variables in the meta-analytic data set that represent the quality aspects or dimensions<br>▪ In case of **multiple quality indicators**, their empirical selection can be informed by the respective correlation matrix and the properties of a measurement model.<br>▪ If multiple indicators are aggregated into quality scores, meta-analysts may choose among several types of **measurement models** (e.g., reflective, causal-formative, or composite factor models). |
| **Step 3:** Study Quality Score Creation | ▪ Creation of **one or multiple scores** representing primary study quality<br>▪ Quality scores can be created *a priori* or as *part of the meta-analytic modeling*.<br>▪ Quality scores can be single or multiple, categorical or continuous, aggregated or separate variables.<br>▪ Quality scores can be extracted from measurement models of multiple quality indicators (e.g., factor scores). |
| **Step 4:** Moderator Analyses | ▪ Estimation of the **moderator effects** of the PSQ score(s) via meta-analytic modeling (e.g., subgroup analyses, mixed-effects meta-regression)<br>▪ Moderator effects may be linear or non-linear. |
| **Step 5:** Moderator Sensitivity Analyses | ▪ Estimation of the **effects of other moderators after controlling for PSQ** (i.e., possible compensatory |

| | |
|---|---|
| | moderator effects) via meta-analytic modeling (e.g., subgroup analyses, mixed-effects meta-regression)<br>▪ Estimation of the **interaction effects between moderators and PSQ** via meta-analytic modeling (e.g., subgroup analyses, mixed-effects meta-regression) |

*Note.* PSQ = Primary study quality.

**Table 2**

*Specification of Selected Meta-Analytic Models to Compare Pooled Effect Sizes and Residual Heterogeneity between High- and Low-Quality Studies*

<table>
<tr><th>Model</th><th>Example specification in the R package “metafor”</th><th>log ℒ</th><th>$AIC_c$</th><th>Within-study variation</th><th>Between-study variation</th></tr>
<tr><td><b>Model 1 (Baseline):</b> Same effect size, same between-study heterogeneity, same within-study heterogeneity</td><td>rma.mv(g, vg,<br>random=list(~1|StudyID/ESID),<br>data=transferct, method=”REML”,<br>tdist=T, test=”t”)</td><td>-563.9</td><td>1133.8</td><td>0.204</td><td>0.281</td></tr>
<tr><td><b>Model 2:</b> Group-specific effect sizes, same between-study residual heterogeneity, same within-study residual heterogeneity</td><td>rma.mv(g, vg,<br>random=list(~1|StudyID/ESID),<br>data=transferct, method=”REML”,<br>tdist=T, test=”t”,<br>mods=~factor(binary.quality))</td><td>-562.1</td><td>1132.2</td><td>0.204</td><td>0.278</td></tr>
<tr><td><b>Model 3:</b> Group-specific effect sizes, group-specific between-study residual heterogeneity, same within-study residual heterogeneity</td><td>rma.mv(g, vg, random=list(~1|ESID,<br>~factor(binary.quality)|StudyID),<br>data=transferct, method=”REML”,<br>tdist=T, test=”t”,<br>mods=~factor(binary.quality),<br>struc=”DIAG”)</td><td>-556.8</td><td>1123.8</td><td>0.213</td><td>Low:<br>0.116<br>High:<br>0.586</td></tr>
<tr><td><b>Model 4:</b> Group-specific effect sizes, same between-study residual heterogeneity, group-specific within-study residual heterogeneity</td><td>rma.mv(g, vg,<br>random=list(~factor(binary.quality)|ESID, ~1|StudyID), data=transferct,<br>method=”REML”, tdist=T, test=”t”,<br>mods=~factor(binary.quality),<br>struc=”DIAG”)</td><td>-558.5</td><td>1127.1</td><td>Low:<br>0.165<br>High:<br>0.302</td><td>0.274</td></tr>
</table>

| | | | | | |
|---|---|---|---|---|---|
| **Model 5:** Group-specific effect sizes, group-specific between-study residual heterogeneity, group-specific within-study residual heterogeneity | `rma.mv(g, vg,`<br>`random=list(~factor(binary.quality)`<br>`\|ESID,`<br>`~factor(binary.quality)\|StudyID),`<br>`data=transferct, method=”REML”,`<br>`tdist=T, test=”t”,`<br>`mods=~factor(binary.quality),`<br>`struc=”DIAG”)` | -554.2 | 1120.5 | Low:<br>0.175<br>High:<br>0.296 | Low:<br>0.130<br>High:<br>0.590 |

*Note.* The variable `binary.quality` is coded as 1 (*high-quality study*) and 0 (*low-quality study*). “Low” and “High” refer to the two categories of study quality. “Group-specific” means that the parameters are estimated for each of the two study-quality groups. The analytic code is based on the R package “metafor”[51]. $\log \mathcal{L}$ = Value of the log-likelihood function, $\mathrm{AIC}_c$ = Corrected Akaike’s Information Criterion.

**Table 3**

*Fit of the Uni- and Multidimensional IRT Models*

| Model | $\log \mathcal{L}$ | $n_p$ | AIC | BIC | CAIC | SRMR | SRMSR |
|---|---|---|---|---|---|---|---|
| Unidimensional 1PL | -112.98 | 7 | 239.96 | 253.75 | 260.75 | 0.138 | 0.175 |
| Two-dimensional 1PL | -112.61 | 9 | 243.23 | 260.96 | 269.96 | 0.138 | 0.167 |
| Unidimensional 2PL | -103.02 | 12 | 230.05 | 253.69 | 265.69 | 0.098 | 0.118 |
| Two-dimensional 2PL | -100.68 | 13 | 227.36 | 252.97 | 265.97 | 0.054 | 0.076 |

*Note.* 1PL = One-parameter logistic (Rasch) model, 2PL = Two-parameter logistic model, $\log \mathcal{L}$ = Value of the log-likelihood function, $n_p$ = Number of parameters, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion, CAIC = Consistent AIC, SRMR = Standardized Root Mean Squared Residual, SRMSR = Standardized Root Mean Square Root of Squared Residuals[71].

**Table 4**

*Mixed-Effects Meta-Regression Model Parameters in the Illustrative Example 3*

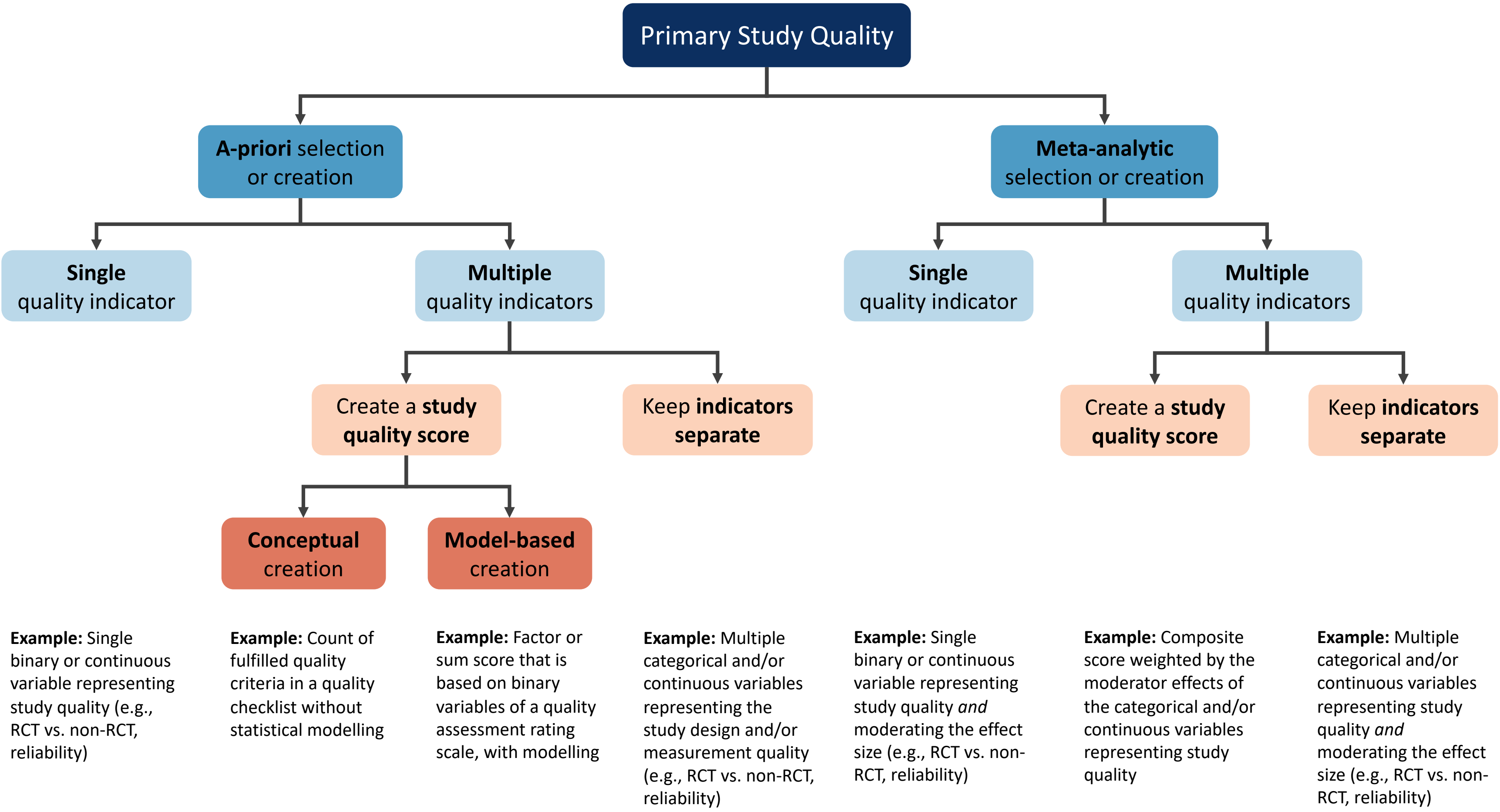
| *Variable* | **Separate quality indicators** | | | **Composite quality score** | | |
|---|---|---|---|---|---|---|
| | *B* | *SE* | *p*-value | *B* | *SE* | *p*-value |
| Intercept | 0.96 | 0.37 | 0.01 | 0.51 | 0.07 | < .001 |
| PubStatus | 0.20 | 0.13 | 0.12 | | | |
| PPCDesign | 0.07 | 0.10 | 0.52 | | | |
| Random | 0.15 | 0.13 | 0.28 | | | |
| TreatedC | -0.50 | 0.08 | < .001 | | | |
| Matched | -0.16 | 0.14 | 0.28 | | | |
| PerfSTA | -0.18 | 0.09 | 0.06 | | | |
| Reliability | -0.49 | 0.43 | 0.26 | | | |
| Composite quality score | | | | 0.28 | 0.05 | < .001 |
| $R_w^2$ | | 0.161 | | | 0.189 | |
| $R_b^2$ | | 0.000 | | | 0.000 | |

*Note.* The regression coefficients were pooled across the 20 imputed data sets. TreatedC = Treated control group(s), PubStatus = Publication Status, Random = Randomization, PPCDesign = Pretest-posttest experimental-control groups design, PerfSTA = Performance-based and standardized assessment, Matched = Matching of the groups, $R_w^2$ = Proportion of the explained variance within studies (i.e., proportional reduction of the within-study variation), $R_b^2$ = Proportion of the explained variance between studies (i.e., proportional reduction of the between-study variation)[72].

Figures

**Figure 1**

*Flowchart of the Decisions on how to Represent Primary Study Quality*


Primary Study Quality
**A-priori** selection or creation
**Meta-analytic** selection or creation
**Single** quality indicator
**Multiple** quality indicators
**Single** quality indicator
**Multiple** quality indicators
Create a **study quality score**
Keep **indicators separate**
Create a **study quality score**
Keep **indicators separate**
**Conceptual** creation
**Model-based** creation

**Example:** Single binary or continuous variable representing study quality (e.g., RCT vs. non-RCT, reliability)

**Example:** Count of fulfilled quality criteria in a quality checklist without statistical modelling

**Example:** Factor or sum score that is based on binary variables of a quality assessment rating scale, with modelling

**Example:** Multiple categorical and/or continuous variables representing the study design and/or measurement quality (e.g., RCT vs. non-RCT, reliability)

**Example:** Single binary or continuous variable representing study quality *and* moderating the effect size (e.g., RCT vs. non-RCT, reliability)

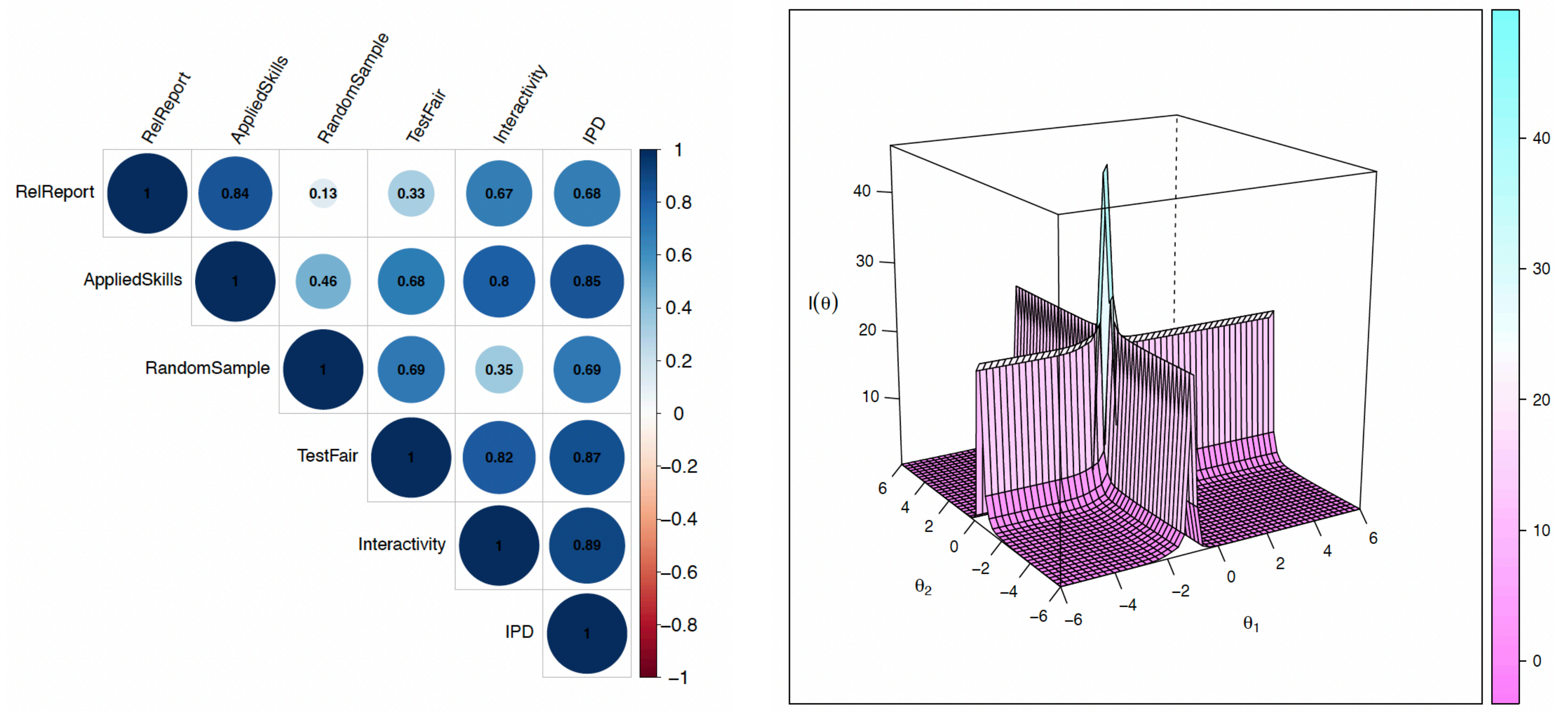
**Example:** Composite score weighted by the moderator effects of the categorical and/or continuous variables representing study quality

**Example:** Multiple categorical and/or continuous variables representing study quality *and* moderating the effect size (e.g., RCT vs. non-RCT, reliability)

*Note.* RCT = Randomized controlled trial.

**Figure 2**

*Plots of (a) the Tetrachoric Correlation Matrix of the Quality Indicators and (b) the Quality Test Information Curve Based on the Two-Dimensional 2PL Model*



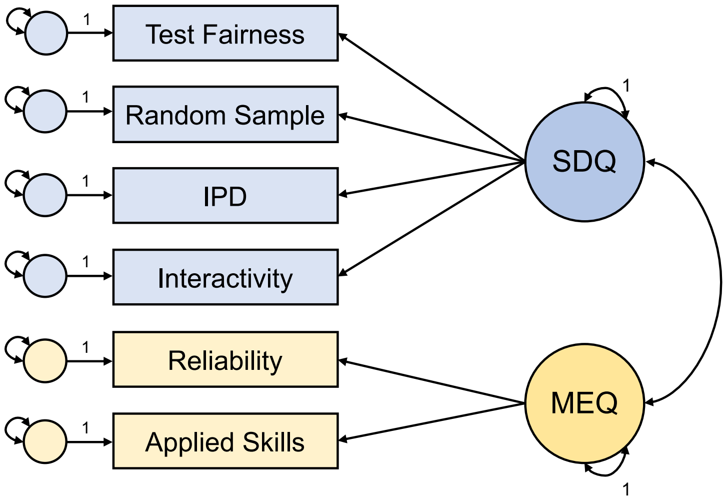
(a) Correlation Matrix

(b) Test Information Curve

*Note.* All indicators were coded binary (*0 = No, the criterion does not apply*; *1 = Yes, the criterion applies*). Panel (a): AppliedSkills = Assessment of applied skills, Interactivity = Task interactivity, IPD = Individual participant data, RandomSample = Random sampling, RelReport = Reliability reporting, TestFair = Test fairness evaluation. Panel (b): The test information plot displays the information function $I(\theta)$ along the two latent quality variables $\theta_1$ and $\theta_2$. Higher values on $I(\theta)$ indicate a more information and better precision on the two latent quality variables $\theta_1$ and $\theta_2$.

**Figure 3**

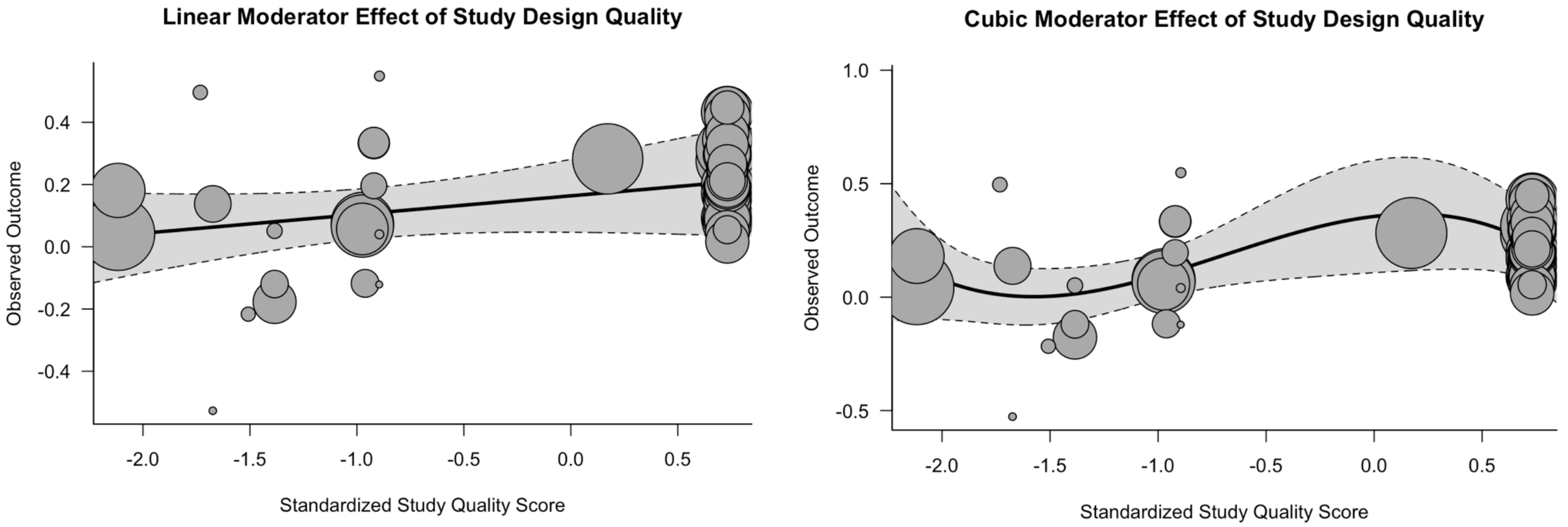
*Reflective Measurement Model of Primary Study Quality in the Illustrative Example 2*



*Note.* IPD = Individual participant data, MEQ = Measurement quality, SDQ = Study design quality. The mean structure is omitted in this path diagram.

**Figure 4**

*Plots of the Linear and Non-Linear Relations between Study Design Quality (Factor 1) and the Effect Sizes in Example 2*

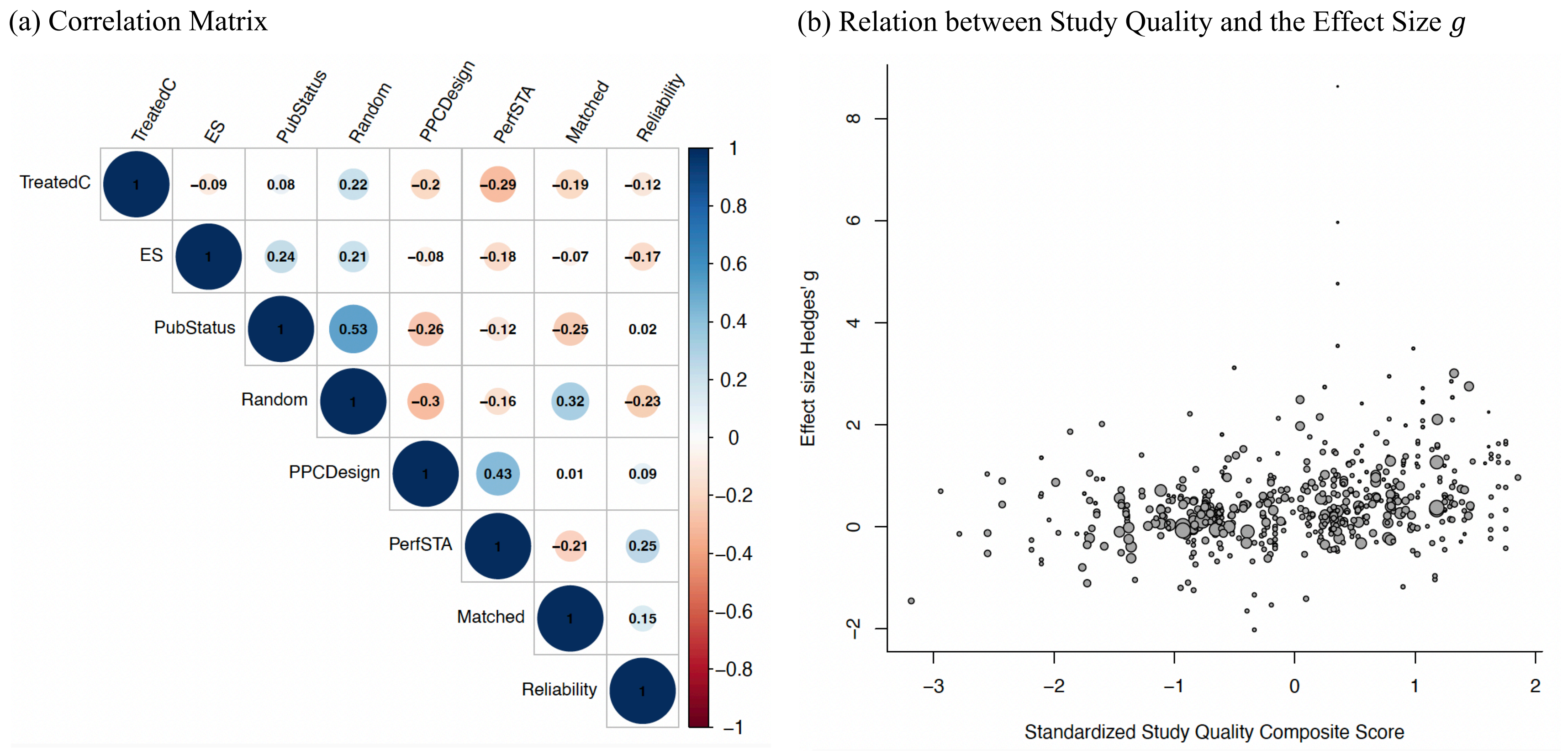


Linear Moderator Effect of Study Design Quality

Cubic Moderator Effect of Study Design Quality

**Figure 5**

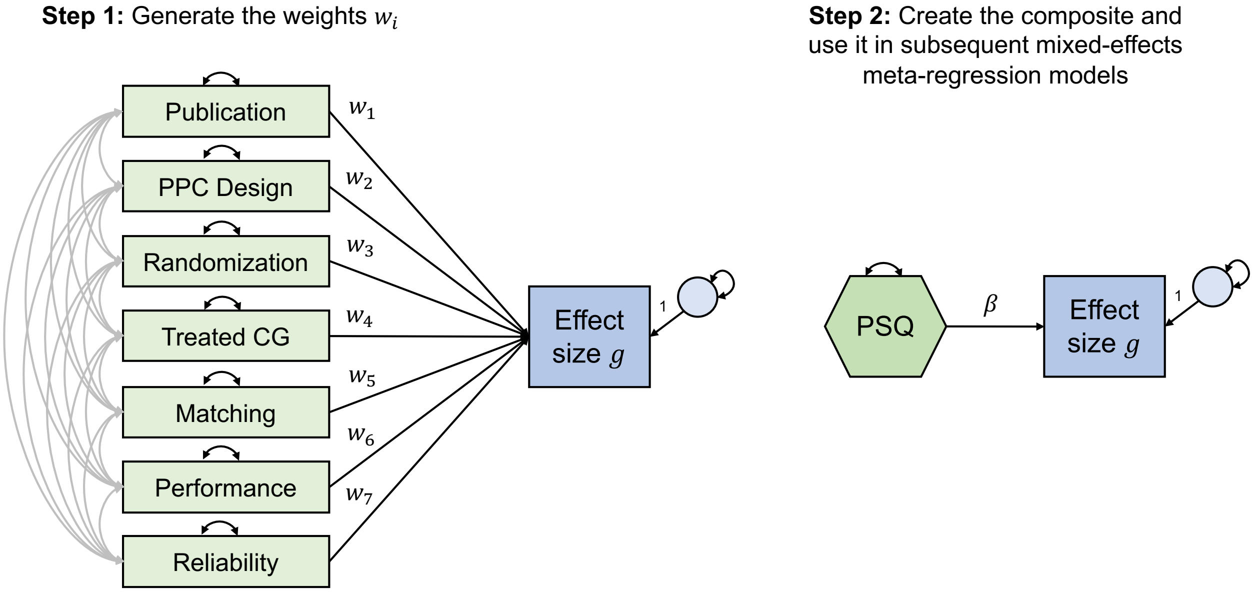
*Plots of (a) the Mixed-Format Correlation Matrix of the Quality Indicators and (b) the Relation between the Primary Study Quality Composite Score and the Effect Size*



*Note.* The correlation matrix was pooled across the 20 imputed data sets. TreatedC = Treated control group, ES = Effect size, PubStatus = Publication Status, Random = Randomization, PPCDesign = Pretest-posttest experimental-control group design, PerfSTA = Performance-based and standardized assessment, Matched = Matching of the groups.

**Figure 6**

*Two-Step Procedure of Creating a Composite Study Quality Score in the Illustrative Example 3*



*Note.* CG = Control group, PPC = Pretest-posttest experimental-control groups, PSQ = Primary study quality. Covariances among the moderators (in grey) inform the model estimation, yet are often not explicit model parameters. The mean structure is omitted in this path diagram.