An Empirical Assessment of Reliable and Valid Measurement as a Prerequisite for Informative Replications

Goos, C.1, Bakker, M.1, Wicherts, J.M.1, & Nuijten, M.B.1

1 Tilburg University

Author note

1 Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, NL.

The authors made the following contributions. CG: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; MB: Conceptualization, Supervision, Writing - Review & Editing; JW: Conceptualization, Supervision, Writing - Review & Editing; MN: Conceptualization, Project Administration, Supervision, Validation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Goos, C., Professor Cobbenhagenlaan 125, 5037 DB, Tilburg, The Netherlands. E-mail: [c.goos@tilburguniversity.edu](mailto:c.goos@tilburguniversity.edu)

Abstract

Recent studies have shown that challenges in measurement practices may have contributed to the “replication crisis” in psychological science. However, the nature and extent of the association between measurement and replicability remains unclear. We investigated the reliability and measurement reporting of 77 measures within 56 Many Labs replications and related original articles [@klein2014investigating; @ebersole2016many; @klein2018many; @ebersole2020many], and their association with replicability. Results from logistic regressions indicated that questionable practices in reporting on measurement (QMPs) in replication studies (b = -5.60, p = .006), but not its reliability (b = 8.21, p =. 106) was associated with lower replicability. Furthermore, Reliability Generalization Meta-analysis revealed that not all measures were sufficiently reliable across contexts. We additionally found that reliability and validity evidence was rarely reported. These findings corroborate existing research that construct validity in published research is low and that may have negative consequences for replications. We offer suggestions on improving measurement practices, and argue that reported measurement information should inform the decision to replicate.

*Keywords:* reliability, validity, measurement, reporting, replicability, credibility, psychology

*Word count:* 159

An Empirical Assessment of Reliable and Valid Measurement as a Prerequisite for Informative Replications

For solid scientific progress in psychology, we need to be able to rely on previous findings - findings should be credible. Unfortunately, psychology seems to face several threats to the credibility of its findings. In this article, we investigate how the credibility of psychological findings may be related to problems in psychological measurement, and in turn, how those problems may affect subsequent replications.

## Credibility & Replicability

A prerequisite for a credible finding is that it is replicable (Vazire et al., 2022). This means that the observed effect should be comparable across the original study and subsequent replications in new samples if the differences between them are believed to be of little relevance to the effect (Nosek et al., 2022).

But replications often fail. In 2015, the Replication Project: Psychology (RPP; Open Science Collaboration, 2015) attempted replications of 100 research findings in psychology, they observed a substantial discrepancy between the replication findings and the original findings. Similarly, the Many Labs projects aimed to replicate the effects reported in original articles across different labs in various locations around the world (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018, 2022), and often found smaller effects in replications compared to original research. Furthermore, the estimates of the effect varied between different lab locations.

In response to this string of failed replications, there was widespread alarm regarding the credibility and robustness of findings in psychological science (Giner-Sorolla, 2019; Hughes, 2018), and efforts to explain the low replication rate . The factors that have been proposed include, questionable research practices (Cumming, 2014; John et al., 2012; Simmons et al., 2011; Wicherts et al., 2016), questionable reporting practices (Bakker & Wicherts, 2011; Nuijten et al., 2016), and biased publication decisions (Bakker et al., 2012; Giner-Sorolla, 2012; Sterling, 1959). Another factor that has gained more traction recently is the impact of measurement-related challenges on replicability and credibility.

## Credibility & Measurement

Psychological constructs, such as affective state or intelligence, cannot be measured directly (Flake et al., 2017). Psychological measurements involve random measurement errors and uncertainties about whether the targeted construct is indeed well reflected in the scores retrieved from the measurement procedure.

The first issue relates to the concept of reliability, which gives an indication of how consistent the responses on a measurement are. Reliability is a critical first step to obtaining credible findings. Because if the scores on a measure are not consistent with themselves, it is unlikely that any effect associated with the measure can credibly be established. Psychometrics offers various ways to assess reliability empirically (Mellenbergh, 2011; e.g. Nunnally, 1978).

The second issue relates to the concept of validity, specifically construct validity, or the overall extent to which a measure measures what it is supposed to measure (Borsboom et al., 2004). Construct validity of test scores is often assessed by comparing the test scores’ empirical associations to the theoretical associations of the constructs (Cronbach & Meehl, 1955). A researcher cannot claim to have found a credible psychological effect, if the measures do not relate to the psychological construct (Cook et al., 2002).

## Replicability & Measurement

While reliable and valid measurement, as well as replicability are important elements that independently contribute to the overall credibility of a finding, there may also be a relation between them. Indeed, recent studies have illustrated that both the reliability of the measure (Stanley & Spence, 2014) and the reporting of information relevant to demonstrating the validity of the measurement (Flake et al., 2017; Flake & Fried, 2020; Shaw et al., 2020) may be related to the chance of successfully replicating a psychological finding. This study aims to investigate the relation of reliability and measurement reporting with replicability, using data and reports from the large scale Many Labs replication projects.

Because the Many Labs projects (REFERENCES) are large-scale collaborations in which multiple lab locations directly replicated the same set of studies, their data allows us to study the variability in theirmeasures across different contexts. Furthermore, as the Many Labs projects used preregistered and documented structured protocols, we therefore believe the measurement use in these projects represent a high standard within the field. Any issues in measurement here might suggest that other replications could also face similar or greater challenges.

## Existing Research

### Reliability.

While reliability is generally considered a measure specific feature, this is false. The reliability of a measure varies between samples (Cho & Kim, 2015; Pauly et al., 2018). Besides random variation due the sampling error, the variations can also reflect true variation in reliability, also known as reliability heterogeneity (Vacha-Haase, 1998). For instance, measures are known to generally show lower reliability in more homogeneous samples as compared to heterogeneous samples (Pike & Hudson, 1998).

The resulting variations in the measurement scores could contribute to the discrepancy between effects in original and replication research, as well as within replications of the same original effect. After all it has long been known that noise in measurement suppresses observed effects and associations (Spearman, 1904). Stanley and Spence (2014) simulated data of items on a scale based on levels of sampling error and reliability similar to what is standard in psychological research. The results showed that the variation due to measurement error was substantial enough to cause replications of a positive small effect to observe anywhere between a medium negative effect and a large positive effect (Stanley & Spence, 2014). In summary, studies using unreliable measures are less replicable, even when the measurement is equally reliable between original and replication.

### Measurement Reporting.

Despite the importance of having reliable and valid measures, research has shown that Questionable Measurement Practices (QMPs) are not uncommon. QMPs are practices that raise doubts about a measurement’s validity (Flake & Fried, 2020), and have been coined as a term analogous to Questionable Research Practices (QRPs, John et al., 2012). QMPs range from lack of transparency and unclear motivation in choice of measure, to poor justification for modifications of a measure and procedure of an existing measure (Flake & Fried, 2020). More specifically, Flake and Fried (2020) identified six key questions than an article should answer to avoid QMPs: the definition of the construct, how the measure was selected, how the measure was operationalized, how the measure was quantified, whether it was modified or not and why, and the reasons and details of creating a measure if applicable. If these questions can be answered based on the reported measurement information, then reporting is sufficiently transparent to promote a more cumulative psychological science.

Flake et al. (2022) documented QMPs among 100 replications and their respective original articles from the RPP (Open Science Collaboration, 2015). They coded the number of measures, the number of items in the measure, and the information that was reported describing the measure and providing evidence justifying its use. Besides limited reporting of validity and reliability evidence for the coded measures, Flake et al. (2022) found that several of the measures in the RPP did not report the number of items, the response format, or the scoring of the scale. Furthermore, only eight of the 40 translated scales contained validity evidence for the translated version of the scale. Measures were similarly modified between original and replication in other ways without evidence showing that the modification did not invalidate the measurement.

Further findings by Flake et al. (2022) and others (Flake et al., 2017; Shaw et al., 2020) also illustrated how QMPs create challenges for replicating researchers. In order to reconstruct the measurement, replication researchers need to know the items that were used, how they were presented to the subjects, and how to compute the scores. If the measure is not reconstructed exactly, constructs measured by original and replication may differ, which may lead to differences in effect size estimates irrespective of the credibility of the original effect.

### Research Contribution.

Our study investigated the relation of reliability and measurement reporting with replicability, using data and reports from the large scale Many Labs replication projects. First, this study expands on the research by Stanley and Spence (2014). Our goal was to see whether reliability was related to replication success in empirical replication data. Specifically, we investigated whether successfully replicated studies differed in reliability from non-replicated studies. Stanley and Spence (2014) assumed reliability to be constant across studies, which we know to be an oversimplification. Therefore, to add further context to the relation between reliability and replication, this study also investigated the variation in reliability as observed across labs.

Furthermore, this research conceptually replicates the study by Flake et al. (2022) on QMPs in replications and original research, now in the context of the Many Labs projects. We then expand on Flake et al. (2022) by exploring associations between QMPs and the replicability of psychological findings.

## Research Questions & Hypotheses

### Reliability.

Measurement reliability varies across samples. Therefore, we investigated what the reliability is in replications of psychological research (RQ1a) and original psychological research (RQ1b).

The Many Labs studies were intended as direct replications. In direct replications, any remaining deviations from the original research should be irrelevant for testing the same effect as the original study (Nosek et al., 2022). As a result, we expected that there would be no difference present in the reliability between replication research and original psychological research (H1).

Reliability has the potential to vary not only between replication and original research, but also across labs assessing the same effect (Vacha-Haase, 1998). We thus investigated whether reliability estimates differed between replicating labs (RQ2).

Heterogeneity in reliability may be contributing strongly to the variation in reliability between labs. However, for effect sizes there is little empirical evidence of widespread heterogeneity among the Many Labs replications (Klein et al., 2018; Olsson-Collentine et al., 2020). We expect that t (H2) as well.

To see if reliability is related to replication success, we investigated whether there is an association between replication study reliability and replication outcome (RQ3).

Greater reliability means the variance around the estimate of the true effect is decreased (Nunnally & Bernstein, 1994). Assuming that the true effect is not null, then the statistical conclusions tested with reliable measures are more likely to converge to significance. Thus, we expected to observe a positive association between reliability in replication research and successful replication of an original finding.

### Measurement Reporting.

This study aims to conceptually replicate the findings on measurement reporting of Flake et al. (2022). To do so we evaluate how common QMPs are in replications of psychological research (RQ4a) and original psychological research (RQ4b).

Flake et al. (2022) found that QMPs were overall more common in the RPP replications as compared to the original research. We hypothesized that for our Many Labs sample QMPs would be more frequent in original psychological research than in replication research (H4), because of the use of structured protocols documenting the measurement.

QMPs obscure information relevant to mimic measurement from original research in subsequent replication attempts. The resulting deviations in measurement may be partly responsible for deviations in replication and original effects (Flake et al., 2022). Thus we investigated whether QMPs in replication protocols were associated with the replication outcome (RQ5).

A reduction in QMPs corresponds to greater transparency in reporting of measurement. In line with earlier research finding other transparency related practices to be associated with more robust estimates of effects (Protzko et al., 2020; Wicherts et al., 2011), we expected QMPs in replications to be negatively associated with replication success (H5).

QMPs obscure information about a particular measurement. This in turn may cause issues for reconstructing that measurement in subsequent replication attempts. We investigated the association between QMPs in original research and QMPs in replications of psychological research (RQ6).

Flake et al. (2022) already proposed that QMPs in original articles may cause issues for recreating measurements for replication research. Shaw et al. (2020) already found such a spill-over effect for validity. In line with this research we expected the total number of QMPs in original psychological research to be positively related to the total number of QMPs in replication research (H6).

# Disclosures

### Preregistration.

We preregistered data collection, coding protocol, and planned analyses: <https://osf.io/jgxyu>. Deviations from the preregistration are explicitly mentioned in the text.

### Data, Materials, and Online Resources.

This manuscript was created in RStudio (Posit team, 2023) with R Version 4.3.1 (R Core Team, 2023), and generated using the Workflow for Open Reproducible Code in Science (WORCS version 0.1.1, Van Lissa et al., 2021) to ensure reproducibility and transparency. All code and data used to generate this manuscript and its results are available at: <https://github.com/CasGoos/measurement_and_replication> and <https://osf.io/9r8yt/>.

### Reporting.

We report how we determined all data exclusions, all manipulations, and all measures in the study. Our sample size was predetermined by the number of studies in the Many Labs projects.

### Ethical Approval.

This research was approved by the Tilburg University School of Social and Behavioral Sciences Ethical Review Board (nr. TSB\_TP\_REMA06).

# Method

## Sample

### Data Source.

The data used for the analyses consisted of three main sources: replication datasets, replication protocols, and original study articles. The data came from the Many Labs replication projects. Specifically, data from Many Labs 1, 2, 3, & 5 (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018). Many Labs 4 (Klein et al., 2022) was excluded, as there was no publicly available replication protocol. Additionally, the replication of (Crosby et al., 2008) in Many Labs 5 made use of videos and eye-tracking measures, which did not match this study’s focus on item-based measures.

### Unit of Analysis.

The unit of analysis in this study was a measure of a single psychological construct used in the primary analysis that was being replicated. Multiple psychological constructs could be measured per original study or replication. We used the the replication protocols to identify these measures. Acquiescence bias checks, manipulation checks, pilot test measures, and measures added for exploratory analyses were not included.

## Data Collection

We retrieved the data on the replication protocols, and replication datasets of Many Labs 1, 2, 3, & 5 from their respective OSF pages: <https://osf.io/wx7ck/>, <https://osf.io/8cd4r/>, <https://osf.io/ct89g/>, & <https://osf.io/7a6rd/>. Both the replication protocols and replication datasets were scanned through to ensure that the planned analyses were feasible. However, no coding or analysis of either of them had taken place before the analyses were preregistered. Further details on the search strategy can be found in the [coding protocol information file](../../SupplementaryMaterials/CodingProtocols/coding_protocol_information.Rmd) in the supplementary materials.

### Replication Datasets.

The replication datasets refer to the publicly available datasets containing the data obtained from all labs that took part in a Many Labs study. For the analyses, we extracted the scores on the items of each previously identified measure that met our inclusion criteria specified below. When scores could not be clearly identified, any available codebooks, analysis scripts or study materials were used to identify the relevant scores.

To be included in the planned analyses on calculated reliabilities, the measure had to be a scale consisting of multiple items. If cleaned data was available this was chosen over raw data, to ensure that variables were coded as intended (e.g. no reverse-coded items). Pilot data were omitted from the analyses. These criteria resulted in suitable item score data from 19 replication sets spread across on average approximately 35 lab locations for the analyses of Hypotheses 2 & 3.

### Replication Protocols.

The replication protocols refer to the publicly available protocols describing the background, methodology, and analysis of each set of replications replicating a single original study across multiple labs. These were retrieved from the OSF pages of the Many Labs projects (the search strategy and OSF file locations can be found in the [Data retrieval information](../../SupplementaryMaterials/data_retrieval_information.Rmd) supplementary document.

### Original Articles.

The original study articles were identified using the citations for these articles in each replication protocol. All articles could be retrieved.

## Measures

We developed and used a coding protocol (the [coding protocol](../../SupplementaryMaterials/CodingProtocols/Measurement_Error_Reporting_Revised_Coding_Protocol.pdf) is available in the supplementary materials) to extract information on the reported measurement evidence and reporting quality for both replication protocols and original studies.

### Measures of Reliability.

Because data from the Many Labs replications were available for each lab location within each replication set, it was possible to calculate reliabilities for any item scale measure used for each lab location separately. This made it possible to assess the variation of reliabilities of the same measure across different contexts.

Cronbach’s Alpha was calculated as the main reliability index used in the analyses, as it remains the most commonly used indicator of the reliability of a measure (Flake et al., 2017), thus allowing for comparisons with the reliabilities reported in original articles. Cronbach’s Alpha was calculated using the alpha function from the psych R package (William Revelle, 2023) with default arguments. Omega was pre-registered as an additional index of reliability, because it is regarded by numerous psychometricians as a more informative alternative to Cronbach’s Alpha (Crutzen & Peters, 2017; Deng & Chan, 2017). The results based on Omega can be found in [Appendix C](AppendixScripts/Appendix_multilevel_analyses.Rmd).

### Measures of Measurement Reporting.

We extracted the reported reliability coefficient and type of index (Cronbach’s Alpha, retest, interrater, etc.) of a measure from both the replication protocols and the original articles when present. Similarly, we coded the presence of any validity evidence, such as a factor analysis, that was presented alongside the measure.

The tests for Hypotheses 4, 5, & 6 were all based on the QMPs coded for both original articles and replication protocols. We included items based on Flake et al. (2022), and additional items to further assess transparent measurement reporting as laid out in Flake and Fried (2020). In total, we coded 20 different QMPs, categorized into five QMP type categories based on Flake and Fried (2020). The QMP categories and example items can be seen in Table 1.

Table   
 *Information of QMP coding variables per category.*

| Category | N Questions | Example Question |
| --- | --- | --- |
| Definition | 1 | A psychological/sociological definition |
|  |  | is given to the name of the measured |
|  |  | variable within the paper. |
| Operationalisation | 5 | The administration format (pen-and- |
|  |  | paper/computer) and environment (in |
|  |  | public/in a lab) are described (Note: |
|  |  | both should be present for a true rating). |
| Selection/Creation | 4 | The source of the scale is provided |
|  |  | (in case the scale was newly developed |
|  |  | this should be clearly stated). |
| Quantification | 4 | The number of items are described. |
| Modification | 6 | Any format changes are mentioned |
|  |  | (paper-and-pencil <–> computer), if no |
|  |  | changes were made to the format, and |
|  |  | this was mentioned then code as No |
|  |  | modification. If it is not clear, then code |
|  |  | as False. |

*Note.* N Questions refers only to the questions used for calculating QMP ratios. Selection and creation share a category as the justifications and requirements in selecting a measure are similar to those for creating a new measure.

QMPs were all coded to be either true, false, or not applicable if not relevant for that measure (e.g. reporting results from a factor analysis for single item measures). For the analyses we calculated a ratio (both per QMP type and overall QMP) based on the number on the number of true responses divided by the number of responses coded to be applicable.

After the initial coding, we made minor revisions to 14 of the 20 QMP items in the preregistered coding protocol, because they were considered too stringent in some of their criteria for what constituted a QMP. For example, in the initial protocol an example item had to be present within the article or protocol itself, or else this was counted as a QMP. In the revised protocol, references to online appendices with example items were also considered sufficient for this item. The analyses, tables and figures presented in this article are all based on the revised coding protocol, the results of the equivalent analyses based on QMPs obtained with our initial protocol can be found in [Appendix D](AppendixScripts/Appendix_initial_QMP_analyses.Rmd).

### Measure of Replication Success.

We used the published reports of the Many Labs projects (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018) to determine replication success based on the reported significance of the meta-analytic effect. An effect was considered successfully replicated if the meta-analytic effect was in the same direction as the original effect and had a p-value lower than .05.

## Analyses

Every hypothesis test in this study was a two-sided test with an alpha of .05. No correction for multiple testing was applied, to ensure that the rate of false negatives remained low. Our hypotheses are about associations that have been proposed in the literature, but to the best of our knowledge have not yet been empirically tested on real data. In this exploratory context, false negatives were considered more detrimental than false positives.

# Results

## Descriptives

Table 2 shows the number of replication sets, measures, and measures per study that were extracted, as well as the proportion of successful replications, for each of the Many Labs separately and in total.

Table   
 *Ratio of measures for which the effect was considered replicated, across many labs projects*

| Many Labs Version | Nr. Measures | Nr. Replicated | Replication Ratio |
| --- | --- | --- | --- |
| 1 | 14 | 12 | 0.86 |
| 2 | 35 | 16 | 0.46 |
| 3 | 15 |  | 0.20 |
| 5 | 13 | 3 | 0.23 |
| Total | 77 | 34 | 0.44 |

*Note.* Nr. measures refers to the total number of primary measures extracted, while Nr. replicated displays the number of measures for which the associated effect was replicated successfully. replication was assessed as unclear for three other measures, since their effect was only partially replicated. These have been treated as not replicated within this table and in further analyses throughout the article.

Our data had a multilevel structure. Each replication was nested within one of four Many Labs projects. Furthermore, 20 replications used more than one primary measure, resulting in 77 measures in both 57 replications protocols and 57 original articles[[1]](#footnote-1). Of these 77 measures 52 were modified from the original to the replication. In the replications, the data of responses on the measures was further nested in several labs per replication. The combined sample size of all labs in the replications was on average 4,844 which was approximately 23 times larger than the average in original studies, which were conducted in single labs.

Because of this multilevel structure we initially planned to conduct multilevel models alongside our single-level regression models for testing Hypothesis 3, 5, and 6. However, these analyses were omitted, because group sizes were small. The results from these analyses are reported in [Supplementary Analyses B](../../SupplementaryMaterials/SupplementaryAnalysesScripts/Supplementary\_multilevel\_analyses.Rmd)) for the sake of completeness and transparency.

## Measurement Reliability

If data from a multiple item scale could be accessed, we calculated the Cronbach’s Alpha of that scale from that data for each lab that administered the scale. As a result, it was possible to include the multiple estimates of Cronbach’s Alpha for those measures into a meta-analysis of the reliability, also commonly referred to as a Reliability Generalization (RG) Meta-Analysis (Botella & Suero, 2012; López-Ibáñez et al., 2024; Vacha-Haase, 1998). This enabled us to quantify the degree of true variation (or heterogeneity) in reliability coefficients across lab locations.

For Cronbach’s alpha, we used formulas 2 & 3 from Duhachek and Lacobucci (2004) to calculate the standard error in the meta-analysis. Heterogeneity was estimated using the tau value, which indicates the standard deviation of the distribution of true Cronbach’s alpha coefficients for a measure, and tested using the Cochran’s Q test for each measure.

The RG meta-analysis was performed using the *rma* function from the *metafor* R package (\*v`r getNamespaceVersion("metafor")[[1]]`\*; Viechtbauer, 2010). Defaults settings were used. We implemented no correction for bias, because the Many Labs replications were not at risk of publication bias.

The average calculated Cronbach’s alpha coefficient across replication sets was 0.787 with a standard deviation of 0.167. Figure 1 displays the distributions of the calculated Cronbach’s Alpha scores from each lab for each measure, separated by successful and unsuccessful replication.

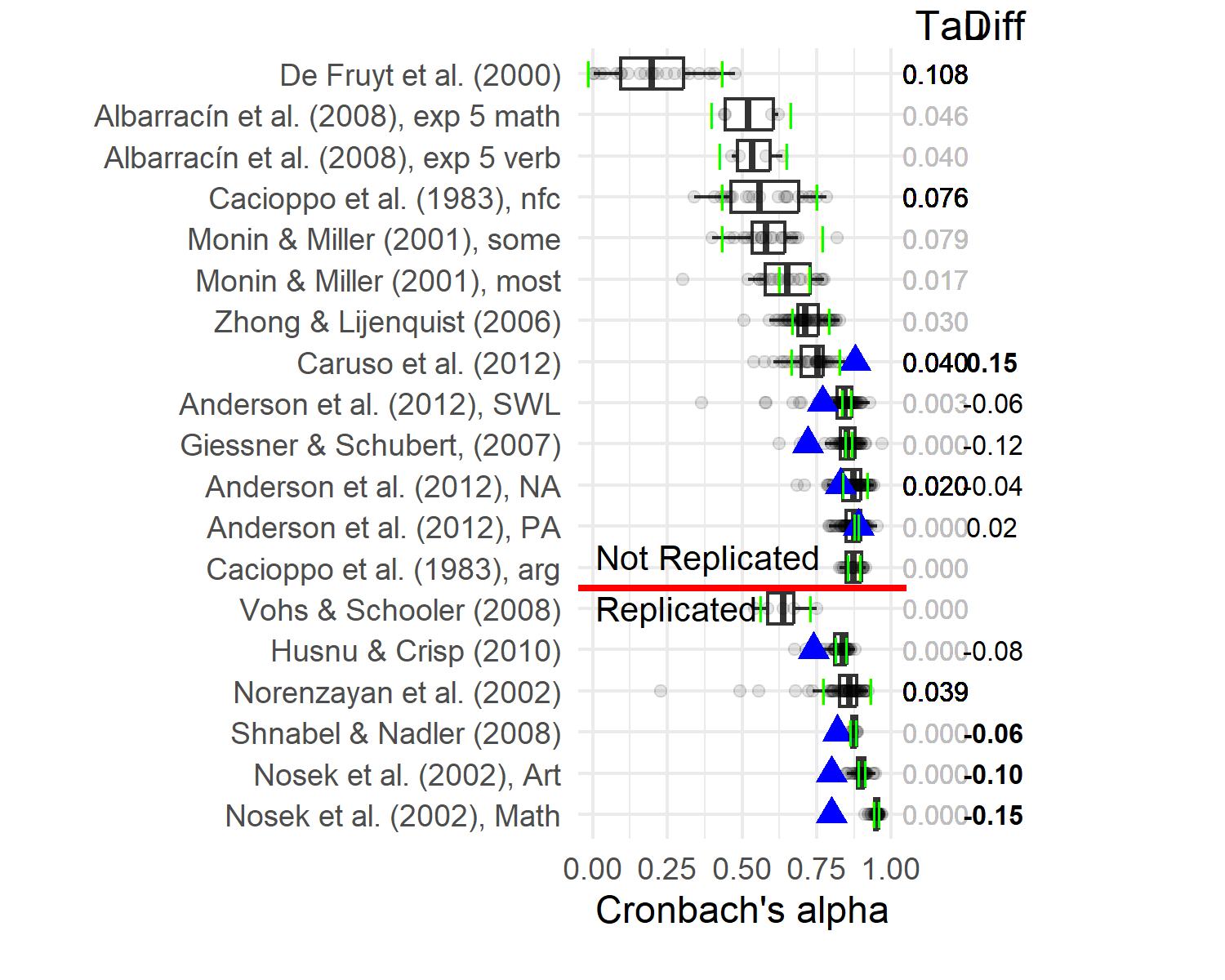


Figure 1: Distributions of calculated Cronbach’s alpha coefficients (> 0) calculated for the responses on a measure at each lab location, across the eighteen distinct measures for which raw data was available from which Cronbach’s alpha coefficients could be calculated. The green lines indicate the meta-analytic prediction interval lower and upper bound. The blue triangles indicate the reported alpha coefficient for that measure from the original article, when reported. The Tau column besides the figure shows the tau heterogeneity estimate based on a meta-analysis of the calculated reliabilities for each measure. Meta-analyses for which the Q-test for heterogeneity was signicant at alpha < .05 are in black, while non-significant results are in grey. The Diff column shows the difference between reported reliability and the average reliability calculated from the Many Labs data for the applicable measures, the reported reliabilities that fell outside the 95% quantile of calculated reliability scores are shown in bold.

The distribution of reliabilities varied across measures. Most of the measures near the bottom showed average reliability scores of at least .80, corresponding to adequate reliability for general research purposes (Nunnally & Bernstein, 1994), with minimal variation between labs. However, other measures showed not only considerably lower average reliability scores, but also greater variation.

Finally, the blue triangles indicate the Cronbach’s Alpha reported in the original articles. The reliabilities were generally only reported for those measures with a large average calculated reliability.

### Hypothesis : Reliability Reporting in Original Compared to Replication Research.

The number of reported reliabilities was 13 in original articles and four in replications, which was too low to ensure that the pre-registered Mann-Whitney U test can function as an informative test on the difference in the reported value of Cronbach’s Alpha between original articles and replications. Instead, the difference between the reported reliability in the original article and the calculated average reliability in replications is displayed in the Diff column in Figure 1. Contrary to Hypothesis 1, this column shows that the reported reliability in original articles was generally lower than the average reliability in the replication sample. Still, more than half of the reported reliabilities did fall within the 95-percentile range around the replication average.

### Hypothesis 2: Within Study Variation in Reliability.

To get an indication of the true variability in the reliability scores, we used the RG meta-analyses to test for heterogeneity in the reliabilities in the replications. This analysis deviated from the multilevel analysis we preregistered, as the preregistered analysis was later deemed not suitable for testing Hypothesis 2 (results from the preregistered analysis are shown in Supplementary Analyses A).

Figure 1 displays the tau estimates, which indicates the differences in true reliability scores between labs for a given measure. For 14 measures the tau estimate was not significantly different from 0, while for five measures, the true variation in reliability across labs was significant. The true reliability was quite variable for some measures. For instance, the estimate of the standard deviation of the true reliability of the top measure in the graph (a measure of conscientiousness from Gosling et al. (2003) used in the replication of De Fruyt et al. (2000)) equaled .046 points of Cronbach’s Alpha.

Thus for around a quarter of measures, the reliability was heterogeneous. Although for the remaining measures, we found no significant heterogeneity in reliabilities, we note that power to detect small heterogeneity is often low (Ioannidis et al. (2007); Olsson-Collentine et al. (2020)).

### Hypothesis 3: Reliability and Replicability.

We used a logistic regression model to test whether replication success as reflected in a significant mean effect in the meta-analysis conducted across the replicating labs could be predicted by the average calculated reliabilities of the measures.

Cronbach’s alpha did not significantly predict replication success in the main logistic regression model (, , 95% , , ). Results based on the Omega coefficient lead to a similar conclusion (see Supplementary Analyses C).

We additionally pre-registered a random intercept and random-slope multilevel regression model of the same relationship with each lab's unique reliability nested within the set of replication of an effect. However, because our dependent variable (replicability) is at the group level, we decided based on suggestions in the literature [@foster-johnson2018Predicting] to use a test on the aggregate level using White's heteroscedasticity adjustment to address the dependency. This analysis also showed a non-significant relation (`r H3\_test\_result\_alpha\_White\_adjusted`). However, it should be noted that for these analyses the reliability coefficient could be calculated for only `r nrow(data\_h3\_avg)` measures. As a result, the estimates of the relation between reliability and replication success obtained using these models each come with large uncertainty.

## Measurement Reporting

### Reliability & Validity Reporting.

Figure 2 depicts the flow of measures in relation to reliability reporting. First, it shows that almost half of the measures in both replication (n = 37) and original research (n = 35) were single-item measures. When looking at the multiple item measures, the graph illustrates that four measures in the replication protocols and 13 in the original articles reported a reliability coefficient. The most commonly reported reliability coefficient was Cronbach’s Alpha, which accounted for three reported reliabilities in the replication protocols, and 11 in the original articles.

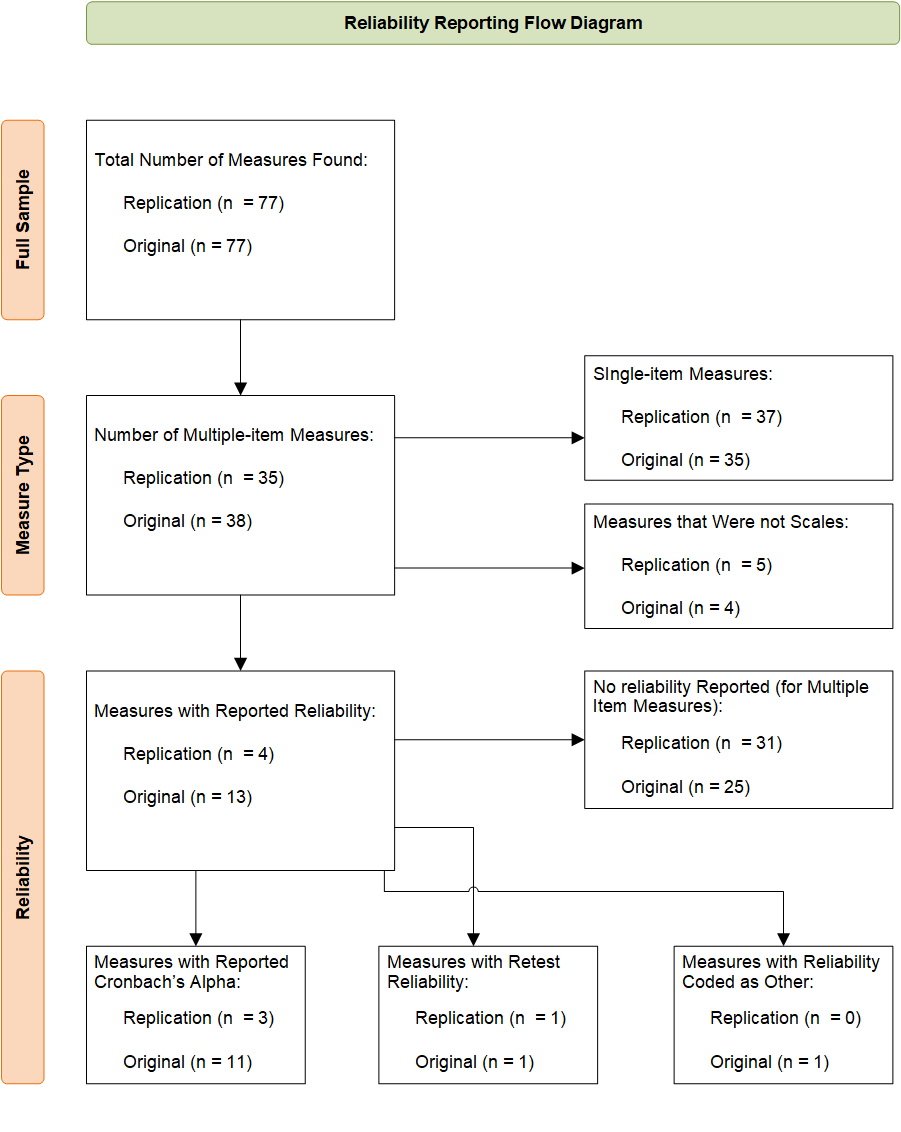


Figure 2: Reliability reporting flow diagram. Figure shows the number of measures as reported in both the replication protocols and original article, which meet the criterion in the box within the diagram and those criteria before it.

Reliability was more commonly reported in original (n =) as compared to replication research (n = 4), (1) = 4.09, *p* = .043. Validity evidence was similarly reported infrequently. eight original articles reported validity evidence from a factor analysis, while only three replications did. Furthermore, we conducted pre-registered exploratory analyses testing the unidimensionality of `r nrow(cfa\_results)` measures with available replication data using a confirmatory factor analysis. We found that `r sum(cfa\_results$Single\_factor == "Yes")` of these measures met our criteria for unidimensionality (see [Supplementary Analyses A](../../SupplementaryMaterials/SupplementaryAnalysesScripts/Supplementary\_pre-reg\_analyses.Rmd) for further details).

### Hypothesis 4: QMPs in Original Compared to Replication Research.

The top part of Figure 3 displays the distributions of the total QMP ratio for both original articles and replication protocols, showing that the average QMP ratio in replication protocols was smaller (0.18) than in original articles (0.31). The bottom of Figure 3 shows that original articles and replication protocols had different distributions of QMPs per category. The number of measures for which modification items were applicable also differed. In total 38.51% of all responses across all QMP items were non-applicable.

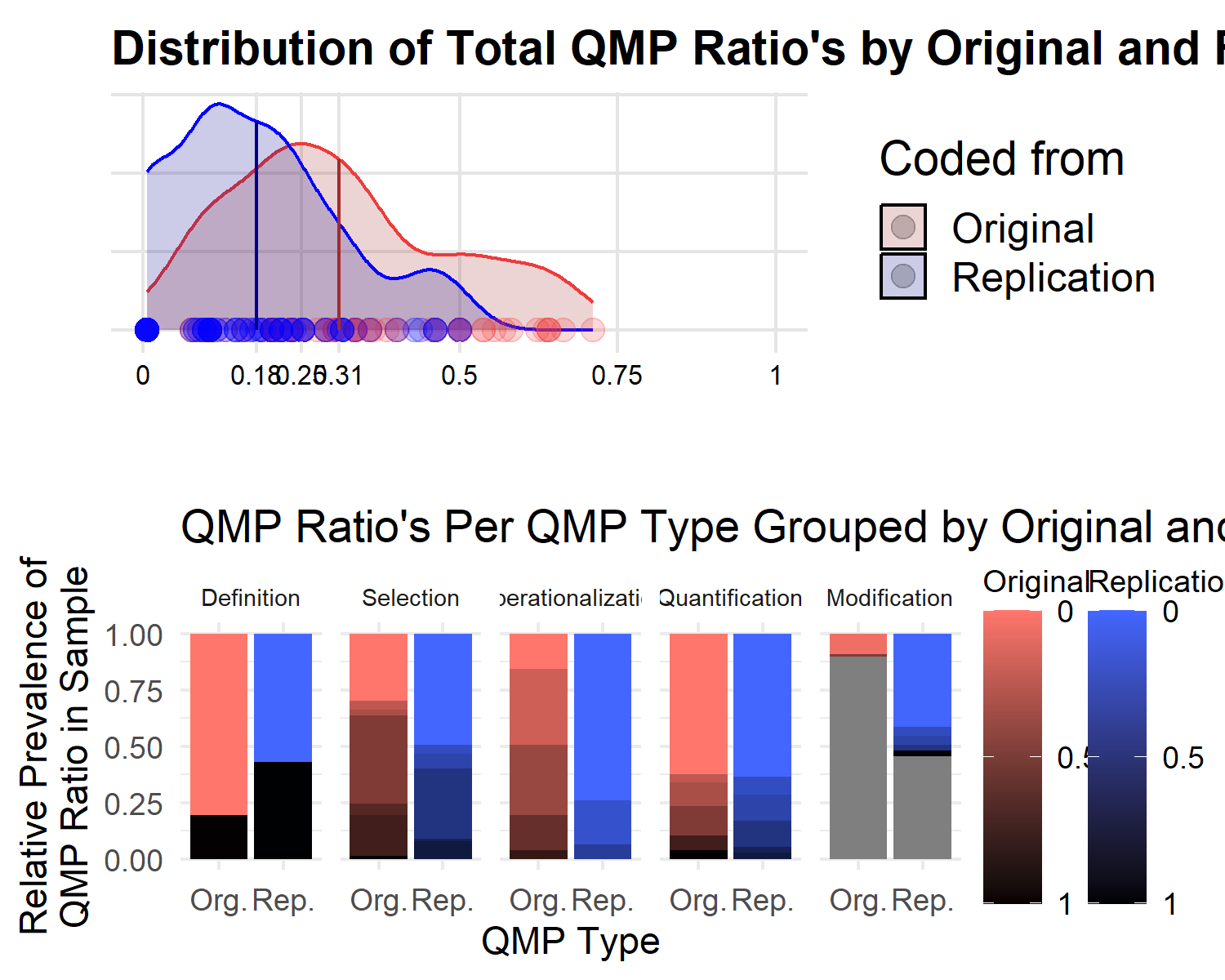


Figure 3: QMP ratio counts for each QMP Type and QMP total ratio distribution grouped by whether the QMP ratio was obtained from an original article or a replication protocol. The top graph shows the distributions of total QMP ratios for both replication protocols and original articles, with the line indicating the mean QMP ratio. The specific observed values are indicated along the bottom row with dots. The bottom row shows for each QMP type the proportions of each QMP ratio obtained, darker colors represent proportionally more QMPs, grey means modification did not occur for that measure.

We used a beta-regression model to test the difference in QMP ratio between original and replication was significant using the *betareg* R package (Cribari-Neto & Zeileis, 2010a; Grün et al., 2012). Beta regression models are similar to other generalized linear regression models, but they are better suited to model dependent variables with values in the interval of , including ratios. Furthermore, they are robust for heteroskedastic and asymmetrically distributed dependent variables (Cribari-Neto & Zeileis, 2010b).

Using a beta-regression model, the total QMP ratio was regressed on a dummy variable indicating whether the coded report was an original article or a replication protocol. The results supported Hypothesis 4 and indicated that this difference was significant (, 95% CI , , ).

### Hypothesis 5: Replication Research QMPs and Replicability.

We used logistic regression to test the association between the QMP ratio in replication protocols and replication success. In line with Hypothesis 5, results showed that successful replications were associated with fewer QMPs compared to unsuccessful replications (, , 95% , , ). However, we did not find the association with the initial protocol in neither the single-level (`r H5\_test\_results\_with\_OR`) nor multilevel models (`r random\_intercept\_h6$full\_result$QMP`)[^2].

[^2FOOTNOTE]: As of writing this article the \*betareg\* package does not allow for estimation of multilevel models. As an approximation, Gaussian random-intercept multilevel models were implemented using the \*lmer\* function from the \*lme4\* package [@R-lme4]. The outcome QMP ratio variable was transformed using the logit link function (the default link function in the \*betareg\* package) to fit the Gaussian model.

### Hypothesis 6: Relation Between QMPs in Original and Replication Research.

Using a beta regression model, we found total QMP ratio in the original article to be significantly related to the total QMP ratio in the subsequent replication (, 95% CI , , ). This result provides evidence in favor of Hypothesis 6. However, this association was sensitive to changes in the QMP protocol, as we did not observe a significant relation between original and replication QMP ratios based on QMPs from the initial protocol (`r betareg\_output\_to\_apa\_full(H6\_test\_results)`). Figure 5 displays this relationship visually.

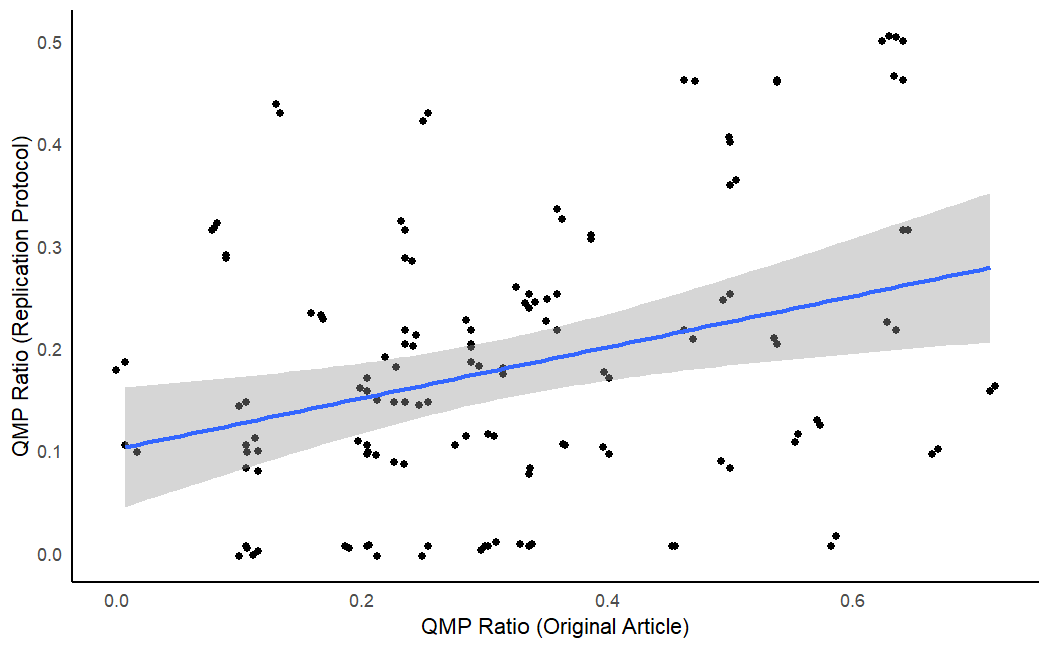


Figure 4: Scatterplot of original QMP ratios and replication total QMP ratios with a linear regression line. The linear regression line is included, as the result based on linear regression was similarly in direction and significance to the beta regression used to test Hypothesis 6 (H6\_glm\_test\_results\_REV) while being easier to understand visually. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots to show the number of observations at points where multiple dots were present.

However, this association was sensitive to changes in the QMP protocol, as we did not observe a significant relation between original and replication QMP ratios based on QMPs from the initial protocol in neither the single-level (`r betareg\_output\_to\_apa\_full(H6\_test\_results)`) nor the multilevel model (`r random\_intercept\_h6$full\_result$QMP`). Furthermore, in the mulitlevel model based on the revised QMPs the association was again not significant (`r random\_intercept\_h6\_REV$full\_result$QMP`).

# We additionally conducted planned exploratory analyses of the association for each unique combination of QMP types, but taking into account the limited sample size, we could not determine any clear relations between types (see Supplementary Analyses A and Supplementary Analyses D for further details on the analyses and results).Discussion

In this article, we analyzed the data, protocols, and related original articles from four Many Labs replication projects to assess reliability and measurement (reporting) practices. We additionally looked at how these features might be related to replication success, as a proxy of the credibility of findings. Overall, even though the average reliability was relatively high, we found that not all measures were sufficiently reliable, nor where measures always equally reliable in each setting in which they were used. Although we did not observe reliability to significantly predict replication success, we did observe a negative relation between replication success and presence of QMPs in replication research. Finally, we found QMPs in the original studies to be positively associated with QMPs in the replications.

## Reliability

### Reliability Varied Within and Between Measures.

While most measures showed sufficient average reliability for most purposes in psychological research (around .80), about a quarter of the measures had a reliability generally considered insufficient (around .60 or less) (Nunnally & Bernstein, 1994). Furthermore, the reliability in original research was generally lower than in the replication samples, contrary to our hypothesis that reported reliability estimates would not differ significantly between replications and original articles. These differences may be due to features that were specific to the Many Labs projects. Large teams of researchers with diverse expertise were involved in the Many labs projects. It is possible that the pooling of expertise and use of structured procedures led to more reliable measurement compared to the original research.

Our results additionally show that the reliability of a measure differs across samples, in some cases even showing signs of true variation, or heterogeneity, in reliability. Reliability heterogeneity was particularly common for measures with a lower average reliability. However, the number of investigated measures was small, and some measures were only used in a small number of independent samples. This is particularly relevant, since the Q-test to test heterogeneity is sensitive to the number of studies included in each test (Li et al., 2015). Furthermore, most measures did not show evidence of significant reliability heterogeneity, a similar lack of widespread heterogeneity was also observed for effect sizes in replications (Klein et al., 2018; Olsson-Collentine et al., 2020). Regardless, our results do demonstrate an often-overlooked aspect of reliability: reliability is a sample-specific, not a measurement-specific feature.

### No Observed Relation Between Reliability and Replicability.

Stanley and Spence (2014) illustrated that measurement error, when not attenuated for, can impact replication assessments. However, we observed no relation between reliability and replicability, even though our chosen replication index did not attenuate for measurement error. This result is surprising not only because it runs counter to the findings of Stanley and Spence (2014) and Hypothesis 3. It also runs counter to the fact that according to statistical theory, the increased noise in the data as the result of lower reliability decreases the effect size estimates based on those data.

However, the relationship between reliability and replicability is complex (LeBreton et al., 2014; Oswald et al., 2015; Willson, 1980; Zhang, 2024). For example, reliability and replicability are only related when the true effect is not null. In our sample, more than half of the effects were not replicated, indicating that many of these effects may be null effects. For these effects no relation between reliability and replicability is expected. Furthermore, while lower reliability decreases effect sizes it can also reduce the observed true variance in the effects (Olsson-Collentine et al., 2023). As a result, effect estimates from various studies will appear more similar than they truly are, seemingly replicating the original. These and other relations are further complicated by the fact that reliability is not a fixed aspect of a scale, but a feature of that scale within a sample (Cho & Kim, 2015; e.g., Vacha-Haase et al., 1999).

The complexity of the relationship combined with the fact that we could only calculate reliability for a small number of measures meant we had low statistical power to detect the overall relation between reliability and replicability, even more so for the multilevel model, which provides another possible explanation for our non-significant findings. NO-ALINEA-BREAK The small number of relevant measures and reported reliabilities also indicates an important issue: we observed a lack of reliability evidence in both original and replication psychological research.

## Measurement Reporting

### Measurement Reporting is Often Incomplete.

The reliability of measures was rarely reported in both original and replication research, which is in line with earlier investigations in the literature (Flake et al., 2017; Flake et al., 2022; Shaw et al., 2020). Reported reliabilities were so few, that the preregistered test for Hypothesis 1 was no longer informative. We found that replication research reported reliability coefficients less often than original research, which is in line with Flake et al. (2022). This pattern continued for validity evidence reporting. Of the 77 measures, validity evidence in the form of factor analysis or similar analyses was reported for eight measures in original articles, and three measures in replication protocols.

The underreporting of reliability and validity evidence has long been discussed (Green et al., 2011; Vacha-Haase et al., 1999; Willson, 1980). Our study illustrates three potential reasons for why this problem has persisted. Firstly, the studies in our sample commonly used single-item measures, an observation that was also made by Shaw et al. (2020). Calculating reliability for single-item measures is not as straightforward and less common as for multiple item measures (Wanous & Reichers, 1996). Concerns with the use of single-item measures go beyond reliability, validity may also be at risk. Single-item measures because they are singular come with an assumption of unidimensionality regarding the construct they, which our exploratory analyses revealed is an assumption that did not frequently hold for multiple-item measures in our sample. However, we did not explore validity concerns unrelated to reporting in-depth, therefore future research is needed to determine the extent of problems with validity.

Secondly, our results indicate that there may be bias in reporting reliability coefficients. Reliabilities were reported in original research for those measures that obtained higher calculated reliabilities in the replication samples. If we take the calculated reliabilities in the replication sample as an accurate representation of the distribution of the true reliability for that measure, then this would imply that reliabilities are more often reported when the measure is truly reliable, and less when it is unreliable; a common issue of selective reporting. Researchers may be reluctant to report reliability for unreliable measures, thus lowering the prevalence of reported reliabilities. In turn, researchers may become overly optimistic about the reliability of measures in their field, similarly to how bias in publication causes issues for establishing estimates of true effect sizes in meta-analyses (Sutton et al., 2000).

Thirdly, the replication protocols were written before data were collected, which means no measurement based on these protocols had taken place yet. As a result, there was no data available yet from which the replicating researchers could have calculated reliability and validity. This would explain why there was little reliability and validity evidence reported in replication protocols. It may still represent a problem however, because original articles similarly rarely reported reliability and validity evidence. This implies that the replications commonly used measures that were not yet validated.

On the other hand, we found that more general issues in the transparency of measurement reporting practices, as assessed by QMPs, were less common in replication protocols compared to original articles. This is in line with Hypothesis 4, but runs counter to the findings of Flake et al. (2022), who found less construct validity evidence in the replications from the RPP (Open Science Collaboration, 2015) than in original articles. We believe this discrepancy was the result of our sample being the Many Labs projects. These projects made use of structured replication protocols to document the way measurement was going to be conducted. For example, several Many Labs replication protocols contained a section specifically listing the deviations from the original measurement.

### Incomplete Measurement Reporting Hinders Replicability.

Existing literature has already warned about the potential detrimental effects of QMPs on replications (Flake et al., 2022; Shaw et al., 2020). It is challenging for replication researchers to mimic the measurement of an original article, when the original article does not document the measurement in sufficient detail. Consequently, the measurement in the replication may assess constructs substantially differently, weakening the relation between the test in original and replication studies (Flake et al., 2022).

In line with Hypotheses 5 & 6, we found an indication of such a spillover effect in our sample. The total QMPs of an original study and the total QMPs in the protocol of the replication for the same study were positively associated. Furthermore, we found that QMPs may have negatively impacted replicability. An increase in total QMP ratio was associated with a decrease in replicability. These associations together indicate that poor measurement reporting in an original study could be a risk factor for subsequent replication attempts.

However, it is worth noting that these associations were only found for QMPs coded with the revised coding protocol. Neither hypothesis was supported by the data on QMPs obtained with the initial coding protocol (see [Appendix D](AppendixScripts/Appendix_initial_QMP_analyses.Rmd) for results based on the initial protocol). Furthermore, these Hypotheses were supported by only one of the multilevel models. However, these models were limited in group and sample size making their interpretation precarious.

## Limitations & Future Research

Our study had several limitations. We encountered numerous challenges in operationalizing QMPs. Whereas previous research has focused on descriptives of specific QMPs detrimental to construct validity (Flake et al., 2022; Shaw et al., 2020), we combined all coded QMPs together as an operationalization to assess measurement practices and evaluate their impact on replicability. However, this also means that it remains uncertain if our operationalization accurately captured the concept. We encountered particular difficulty in determining when a practice is questionable, and the impact of the context on this decision. As a result, we already deemed it necessary to revise our coding protocol to be more lenient regarding more context dependent QMPs compared to the initial protocol. This revision was not trivial, as it altered the interpretation of our results. This finding reflects the challenges we encountered in operationalizing QMPs in a single index.

Another challenge in constructing a QMP variable was the way to treat practices that were not applicable to a particular measurement. We chose to construct ratios to cancel out the effects of non-applicable items as much as possible. However, this has the consequence that both a measure for which all relevant measurement-related information was reported, and a measure for which no item was applicable would have a QMP ratio of zero. Yet, we would not consider the latter case to indicate the same level of completeness as the first. A little more than one third of all responses were not applicable. As a result, the relation between QMP ratios and measurement reporting completeness may have been obscured to some extent.

Interpreting any relation between QMP ratios and measurement reporting is complex, as it is not clear how high a ratio of QMPs should be to represent clear violations of good research practices. One might argue that any QMP is a sign of questionable research and thus that any ratio greater than zero is cause for alarm. However, this may be too harsh. As mentioned before, not all violations may have detrimental consequences or be relevant in all contexts. Future research will be needed to determine the best way(s) to operationalize QMP ratios. Researchers can take inspiration from this study and earlier studies by Shaw et al. (2020) and Flake et al. (2022), as well as the criteria on measurement reporting as set out in the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.) (2014).

A second limitation is that our tests on reliability were likely underpowered. This was the result of the small number of reported reliability scores, and multiple item scales with available data. However, this limitation also illustrates two key findings. One finding is that reliability was rarely reported for measures in both original articles and replication protocols.

The other finding is that around half of the measures in our sample were single item-measures. The use of single item measures comes with psychometric risks (Diamantopoulos et al., 2012; Nunnally, 1978), and is more limited with respect to the type of reliability and validity evidence that can be determined (Sarstedt et al., 2016; Shaw et al., 2020). Because of these issues, it could even be argued that the use of single-item measures is a QMP in certain contexts.

As a third limitation, it is important to consider whether replication protocols and research articles can be fairly compared in terms of QMPs. The space for describing methodology within a replication protocol is generally shorter than an article. Furthermore, for direct replications the methodology may be seen as largely equivalent to the original article, thus measurement descriptions in the replication protocol may be considered superfluous. There are three main reasons why we believe protocols and articles remain comparable. Firstly, the majority of the measurements in our replication sample were modified in some way, it is therefore important that replication protocols still report on their measurement as it may deviate from the original. Secondly, articles are often also restricted in the amount of space they have available to devote to measurement, so space restrictions are no reason per se to not report important measurement information (Gardiner, 2019). Thirdly, supplementary materials were also included as a valid information source, which increase the space available to describe the measurement, further equalizing reporting space in original articles and replication protocols. Still, future research may wish to analyze supplementary materials in greater detail, to investigate to what extent these materials cover measurement details in both replication protocols and original articles.

A fourth limitation is that we of only used one definition of replication success. Replication success can be estimated and framed in multiple ways (for an example in replication research, see Open Science Collaboration, 2015), including methods that do not make use of significance testing, or that attenuate for (measurement) error. The relationship between reliability and other indices of replication success may differ.

Finally, we note that we used an observational design which means that we cannot attach definitive causal directions to the relations observed in this study. To directly assess if there is a negative impact of reliability and QMPs on replicability, we would need experimental research. Future research could for example randomly assign original studies in a large scale replication project to either needing to report all relevant measurement information as noted down in Table 1 in Flake and Fried (2020), while other studies receive no such reporting requirements. Similarly, the requirement to use validated measures based on the criteria set out in section 1, 2, and 3 in the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.) (2014) standards, or lack thereof could be randomly assigned to conditions. Although such experimental designs would be the most direct way to assess causality, it would take tremendous coordination and funding to integrate original studies into a large scale replication project from the start. While such efforts are not unprecedented (Protzko et al., 2020), alternative methods also exist to assess causal relations on observational data, at least to some extent. Rohrer (2018) has suggested researchers make use of Directed Acyclic Graphs to investigate causality in (non-)experimental settings by systematically accounting for confounding influences on an effect.

## Recommendations

Taking all the findings together, we formulated several recommendations. Firstly, reliable and valid measures are a prerequisite for credible findings. The observation that many measures are not reported with such evidence is worrying, because it obscures the credibility of findings, and hampers scientific progress. Therefore, we join earlier recommendations (Barry et al., 2014; Flake et al., 2022; Flake & Fried, 2020; Hogan & Agnello, 2004) that establishing validity evidence for multiple item scales should become a community standard for credible research. Furthermore, reliability coefficient estimates for multiple item scales should by default always be reported, since these estimates can easily be obtained with most statistical software, and provide important information on a measure’s performance. Only researchers that can provide reasoning to the contrary for their research should be exempt from this default.

Secondly, researchers should report their procedure and measurement in detail, as specified in for example Table 1 in Flake and Fried (2020), or section 3.6 in American Psychological Association (2020) on measures and measurement. This is important so that it can be mimicked in replications if needed. The remaining information that could not fit in the article can be made available on a public repository, such as the Open Science Framework (OSF; Soderberg (2018)) or Zenodo (“Zenodo Open Data Repository (CERN),” n.d.).

Furthermore, we argue that researchers seeking to replicate a study should first evaluate the measurement of that study. It is crucial for an informative replication that the measurement is reliable or valid, otherwise it may be futile to attempt a direct replication (see also Nuijten, 2022). If the measurement is insufficiently reliable or valid, we suggest that researchers instead use their resources to conduct a replication of a study with reliable, valid and well-documented measurement. When replicating another study is not an option, we advise the replicating researcher to first attempt a conceptual replication using reliable and valid measurement, or use both the original and their own preferred measure (as in Stoevenbelt et al., 2018). Afterwards, a direct replication can be performed based on the conceptual replication to further assess the effect’s credibility.

# Conclusion

Through our investigations into the reliability and reporting of measurement in the Many Labs replications and associated original studies, we found that reliability and validity evidence was often not reported. Furthermore, QMPs that obscured important information needed to evaluate and reconstruct the measurement were common, especially in original research. We observed results suggesting that QMPs in turn related to the replicability of a finding. Results were less clear for reliability differences between original and replication, most likely because the analysis lacked power to detect a difference, which was a concern here even more strongly than for the other analyses. However, we did observe that around a quarter of measures showed signs of significant variation in reliability across labs. Combined these findings illustrate the need for widespread adoption of measurement reporting standards, and that researchers should consider the validity and reporting completeness of a study’s measurement before attempting to replicate it.

## ORCID iDs

Cas Goos <https://orcid.org/0009-0005-3792-4148> Marjan Bakker <https://orcid.org/0000-0001-9024-337X> Jelte M. Wicherts <https://orcid.org/0000-0003-2415-2933> Michèle B. Nuijten <https://orcid.org/0000-0002-1468-8585>

## Funding

The preparation of this article was supported by the veni grant ??? awarded to Michèle Nuijten, and the vici grant ??? awarded to Jelte M. Wicherts from the Netherlands Organization for Scientific Research (NWO).

# References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Publication manual of the american psychological association 2020: The official guide to APA style*. American Educational Research Association.

American Psychological Association. (2020). *Publication manual of the american psychological association 2020: The official guide to APA style* (7th ed.). American Psychological Association.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior: A Review of Seven Journals. *Health Education & Behavior*, *41*(1), 12–18. <https://doi.org/10.1177/1090198113483139>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Botella, J., & Suero, M. (2012). Managing Heterogeneity of Variance in Studies of Reliability Generalization With Alpha Coefficients. *Methodology*.

Cho, E., & Kim, S. (2015). Cronbach’s coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Cribari-Neto, F., & Zeileis, A. (2010a). Beta regression in R. *Journal of Statistical Software*, *34*(2), 1–24. <https://doi.org/10.18637/jss.v034.i02>

Cribari-Neto, F., & Zeileis, A. (2010b). Beta regression in *r*. *Journal of Statistical Software*, *34*(2). <https://doi.org/10.18637/jss.v034.i02>

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. <https://doi.org/10.1037/h0040957>

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, *19*(3), 226–228.

Crutzen, R., & Peters, G.-J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*(3), 242–247.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. <https://doi.org/10.1177/0956797613504966>

De Fruyt, F., Van De Wiele, L., & Van Heeringen, C. (2000). Cloninger’s Psychobiological Model of Temperament and Character and the Five-Factor Model of Personality. *Personality and Individual Differences*, *29*(3), 441–452. <https://doi.org/10.1016/S0191-8869(99)00204-4>

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement*, *77*(2), 185–203.

Diamantopoulos, A., Sarstedt, M., Fuchs, M., Wilczynski, P., & Kaiser, S. (2012). Guidelines for Choosing Between Multi-Item and Single-Item Scales for Construct Measurement: A Predictive Validity Perspective. *Journal of the Academy of Marketing Science*, *40*, 434–449. <https://doi.org/10.1007/s11747-011-0300-3>

Duhachek, A., & Lacobucci, D. (2004). Alpha’s standard error (ASE): An accurate and precise confidence interval estimate. *The Journal of Applied Psychology*, *89*(5), 792–808. <https://doi.org/10.1037/0021-9010.89.5.792>

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. <https://doi.org/10.1177/2515245920958687>

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576–588. <https://doi.org/10.1037/amp0001006>

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. <https://doi.org/10.1177/1948550617693063>

Gardiner, E. (2019). Editorial: Methods papers. *Platelets*, *30*(1), 2–2. <https://doi.org/10.1080/09537104.2018.1529865>

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562–571.

Giner-Sorolla, R. (2019). From crisis of evidence to a “crisis” of relevance? Incentive-based answers for social psychology’s perennial relevance worries. *European Review of Social Psychology*, *30*(1), 1–38.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. <https://doi.org/10.1016/S0092-6566(03)00046-1>

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in psychological assessment: Recognizing the people behind the data. *Psychological Assessment*, *23*(3), 656.

Grün, B., Kosmidis, I., & Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, *48*(11), 1–25. <https://doi.org/10.18637/jss.v048.i11>

Hogan, T. P., & Agnello, J. (2004). An Empirical Study of Reporting Practices Concerning Measurement Validity. *Educational and Psychological Measurement*, *64*(5), 802–812. <https://doi.org/10.1177/0013164404264120>

Hughes, B. (2018). *Psychology in crisis*.

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ (Clinical Research Ed.)*, *335*(7626), 914–916. <https://doi.org/10.1136/bmj.39343.408449.80>

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., … Ratliff, K. A. (2022). Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), 35271. <https://doi.org/10.1525/collabra.35271>

Klein, R. A., Ratliff, K., Vianello, M., Adams Jr, R. B., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., & Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, *45*(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahnìk, Š. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. <https://doi.org/10.1177/2515245918810225>

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for Criterion Reliability in Validity Generalization: A False Prophet in a Land of Suspended Judgment. *Industrial and Organizational Psychology*, *7*(4), 478–500. <https://doi.org/10.1017/S1754942600006775>

Li, S., Jiang, H., Yang, H., Chen, W., Peng, J., Sun, M., Lu, C. D., Peng, X., & Zeng, J. (2015). The Dilemma of Heterogeneity Tests in Meta-Analysis: A Challenge from a Simulation Study. *PLoS ONE*, *10*(5), e0127538. <https://doi.org/10.1371/journal.pone.0127538>

López-Ibáñez, C., López-Nicolás, R., Blázquez-Rincón, D. M., & Sánchez-Meca, J. (2024). Reliability generalization meta-analysis: Comparing different statistical methods. *Current Psychology*. <https://doi.org/10.1007/s12144-023-05604-y>

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. Eleven international publishing.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>

Nuijten, M. B. (2022). Assessing and improving robustness of psychological research findings in four steps. In *Avoiding questionable research practices in applied psychology* (pp. 379–400). Springer.

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226.

Nunnally, J. C. (1978). *An Overview of Psychological Measurement* (B. B. Wolman, Ed.; pp. 97–146). Springer US. <https://doi.org/10.1007/978-1-4684-2490-4_4>

Nunnally, J. C., & Bernstein, I. H. (1994). The assessment of reliability. *Psychometric Theory*, *3*, 248–292.

Olsson-Collentine, A., Bakker, M., & Wicherts, J. (2023). *Unreliable Heterogeneity: How Measurement Error Obscures Heterogeneity in Meta-analyses in Psychology*. <https://doi.org/10.31234/osf.io/jnb6e>

Olsson-Collentine, A., Wicherts, J. M., & Assen, M. A. L. M. van. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*, 922–940. <https://doi.org/10.1037/bul0000294>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

Oswald, F. L., Ercan, S., McAbee, S. T., Ock, J., & Shaw, A. (2015). Imperfect Corrections or Correct Imperfections? Psychometric Corrections in Meta-Analysis. *Industrial and Organizational Psychology*, *8*(2), e1–e4. <https://doi.org/10.1017/iop.2015.17>

Pauly, M., Umlauft, M., & Ünlü, A. (2018). Resampling-based inference methods for comparing two coefficients alpha. *Psychometrika*, *83*(1), 203–222. <https://doi.org/10.1007/s11336-017-9601-x>

Pike, C. K., & Hudson, W. W. (1998). Reliability and Measurement Error in the Presence of Homogeneity. *Journal of Social Service Research*, *24*(1-2), 149–163. <https://doi.org/10.1300/J079v24n01_07>

Posit team. (2023). *RStudio: Integrated development environment for r*. Posit Software, PBC. <http://www.posit.co/>

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., & Lundmark, S. (2020). *High replicability of newly-discovered social-behavioral findings is achievable*.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. <https://doi.org/10.1177/2515245917745629>

Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016). Selecting single items to measure doubly concrete constructs: A cautionary tale. *Journal of Business Research*, *69*(8), 3159–3167. <https://doi.org/10.1016/j.jbusres.2015.12.004>

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from many labs 2. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 289.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Soderberg, C. (2018). Using OSF to Share Data: A Step-by-Step Guide. *Advances in Methods and Practices in Psychological Science*, *1*, 251524591875768. <https://doi.org/10.1177/2515245918757689>

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. <https://doi.org/10.2307/1412159>

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. <https://doi.org/10.1177/1745691614528518>

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Stoevenbelt, A. H., Wicherts, J., Flore, P., Torka, A.-K., & Gomez, L. A. (2018). *RRR: Johns, Schmader and Martens (2005)*. <https://doi.org/10.17605/OSF.IO/ZWNXY>

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *BMJ (Clinical Research Ed.)*, *320*(7249), 1574–1577. <https://doi.org/10.1136/bmj.320.7249.1574>

Vacha-Haase, T. (1998). Reliability Generalization: Exploring Variance in Measurement Error Affecting Score Reliability Across Studies. *Educational and Psychological Measurement*, *58*(1), 6–20. <https://doi.org/10.1177/0013164498058001002>

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education*, *67*(4), 335–341.

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A.-L., Peikert, A., Struiksma, M. E., & Vreede, B. M. I. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, *4*(1), 29–49. <https://doi.org/10.3233/DS-210031>

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility Beyond Replicability: Improving the Four Validities in Psychological Science. *Current Directions in Psychological Science*, *31*(2), 162–168. <https://doi.org/10.1177/09637214211067779>

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports*, *78*(2), 631–634.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WISC-IV and clinical validation of the four-and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*(2), 114–131.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, *6*(11), e26828.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832.

William Revelle. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. <https://CRAN.R-project.org/package=psych>

Willson, victor L. (1980). Research Techniques in AERJ Articles: 1969 to 1978. *Educational Researcher*, *9*(6), 5–10. <https://doi.org/10.3102/0013189X009006005>

Zenodo open data repository (CERN). (n.d.). In *European University Institute*. Retrieved May 23, 2024, from [https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/ DataPortal/Zenodo](https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/%20%20%20DataPortal/Zenodo)

Zhang, Q. (2024). Meta-analysis of correlation coefficients: A cautionary tale on treating measurement error. *Psychological Methods*, *29*(2), 308–330. <https://doi.org/10.1037/met0000498>

# Supplementary Analyses A: Pre-registration Analyses

This appendix contains the pre-registered analyses that were either changed or omitted from the article.

## Intraclasscorelation Assessment for Within Measure Variation in Reliability

### Pre-registration Analysis Justification & Reason for Omission from Article.

Reliabilities could be calculated from the replication datasets. On top of that, reliabilities could be calculated for each individual lab per replication set so that the variance of them could be assessed. To separate between and within variance in reliability, a multilevel random intercept model was specified with replication set as the grouping variable and the Cronbach’s alpha or omega value of a replication as dependent using the *lmer* function from the *lme4* package [\*v`r getNamespaceVersion("lme4")[[1]]`\*; @R-lme4].

The purpose of this analysis was to investigate variability in calculated reliability coefficients within and between replication sets. However, at the time of pre-registering we did not know of any formal test for the significance of within group variance after controlling for between group variance. Instead, the intraclass correlation (ICC) coefficient was calculated an indicator of within vs. between variability.

This analysis was omitted from the article because it did not link well enough to Hypothesis 2. The test results also do not give use a very informative representations of variance in reliability in the data for us to draw further inferences from. For instance, it assumes that we should observe large differences in reliabilities between measures for this to provide an accurate test of whether the variance within measures is large or not.

### Results.

The between-group and within-group variance for Cronbach’s alpha were approximately 0.033 and 0.008 respectively. The resulting ICC was approximately 0.805. This suggests that variance was larger between replication sets than within. These results are generally in line with hypothesis 2. However, this statistic does not necessarily give a test of hypothesis 2, as outlined above.

## Unidimensionality Exploratory Analyses

### Pre-registration Analysis Justification & Reason for Omission from Article.

To present additional validity evidence alongside the analyses of Cronbach’s alpha and omega coefficients, we investigated the unidimensionality of each measure. To test this, each multiple item scalar measure had their factor structure investigated on the entire measure data of the replication set.

This analysis was omitted from the main article since determining when a measure should be unidimensional would require an in-depth review and expert evaluation for the literature surrounding the measure, which is outside of the scope of this article and the authors’ expertise. Additionally, this information may also simply not be available, further complicating the analyses.

For the sake of this analysis the simple heuristic was chosen that if the measure was used as a single composite variable in the analyses, then it should be unidimensional. This is somewhat similar to what was done by Shaw et al. (2020), but for the sake of this article it was deemed to strong an assumption to make that this heuristic would always hold. For instance, IQ score is often used as a single indicator score, however the construct of IQ is not considered unidimensional, and is instead understood to have many sub-dimensions (e.g. verbal comprehension, perceptual reasoning, etc.). As a result, the use of a single score in the analyses cannot be taken as evidence that the construct is considered unidimensional by neither the authors, nor the broader literature on the topic. NOALINEABREAK Nevertheless, we present here the result of the analyses made under the assumption that the one score, one dimension heuristic holds.

## Model

A single factor confirmatory factor analysis was implemented using the *fa* function from the *psych* VERSION package [\*v`r getNamespaceVersion("psych")[[1]]`\*; @R-psych]in R. The maximum likelihood factoring method was used, with defaults being used for all other function arguments. From this analysis, the RMSEA of the one factor solution was extracted. Additionally, a parallel test was conducted using the *fa.parallel* function also from the *psych* package. If either the RMSEA score < .08, or the parallel analysis returned a one factor solution, unidimensionality was coded as true, otherwise it was coded as false.

## Results

Suitable data for factor analyses was available for nineteen measures. These measures were the same as those shown in Figure 1. The results are shown in Table 3. A composite score based on the measurement responses was used for sixteen of the nineteen measures to form a single variable in the analyses and were thus regarded as intended to be unidimensional. The number of dimensions for the remaining three were unclear. Sufficient fit of the unidimensional model was found for seven out of the nineteen evaluated measures. However, as shown in Table 3, based on the parallel analysis test alone the unidimensional model held for only two of the nineteen measures. For those two measures and an additional five the RMSEA criteria suggested evidence for single factor structures.

Table   
 *RMSEA and Parallel Analysis Suggested Factors based on Factor Analysis for Measure Data*

| Original Article | RMSEA | N Factors | Unidimensional |
| --- | --- | --- | --- |
| Caruso et al. (2012) | 0.116 | 3 | No |
| Husnu & Crisp (2010) | 0.140 | 2 | No |
| Nosek et al. (2002), Art | 0.149 | 3 | No |
| Nosek et al. (2002), Math | 0.136 | 2 | No |
| Anderson et al. (2012), SWL | 0.036 | 2 | Yes |
| Anderson et al. (2012), PA | 0.097 | 4 | No |
| Anderson et al. (2012), NA | 0.104 | 5 | No |
| Giessner & Schubert (2007) | 0.192 | 3 | No |
| Norenzayan et al. (2002) | 0.083 | 10 | No |
| Zhong & Liljenquist (2006) | 0.177 | 2 | No |
| Monin & Miller (2001), most | 0.042 | 2 | Yes |
| Monin & Miller (2001), some | 0.062 | 3 | Yes |
| Cacioppo et al. (1983), elm | 0.077 | 2 | Yes |
| Cacioppo et al. (1983), nfc | 0.062 | 3 | Yes |
| De Fruyt et al. (2000), consc | Did | Not | Converge |
| Albarracín et al. (2008), exp. 5, math | 0.016 | 1 | Yes |
| Albarracín et al. (2008), exp. 5, verb | 0.000 | 1 | Yes |
| Shnabel & Nadler (2008) | 0.182 | 4 | No |
| Vohs & Schooler (2008) | 0.112 | 2 | No |

*Note.* N Factors refer to the number of factors suggested based on parallel analysis of the measure data. The word or abbreviation after the author names and year indicates the specific measure within that original article that was analyed.

These results are potentially quite worrying given that many studies made use of single item measures, which assume unidimensionality by default without being able to assess it. These results show that for the measures that could be tested for unidimensionality, and that were mostly used as single indicators in the analyses for the replication protocol, unidimensionality commonly did not hold. However, one could argue that measures being used as singular indicators does not necessarily mean these measures all assessed constructs with a unidimensional structure. For example, IQ is often used as a singular indicator, even though it consists of several factors (Weiss et al., 2013). However, it should be noted that at least with respect to Many Labs 2, all measures that were considered unidimensional for the analyses in this study were also considered unidimensional in the investigation by Shaw et al. (2020) who also observed a lack of unidimensionality evidence in their investigation of Many Labs 2. To conclude, there is an indication that validity evidence for measures used in replication research is lacking, and that this may have further implications beyond the measures for which it could be assessed here.

## H6 QMP-Category Specific Exploratory Analyses

### Pre-registration Analysis Justification & Reason for Omission from Article.

Follow-up analyses were preregistered for hypothesis 6. The pre-registered plan was to make use of contingency tables to compare relations between all five QMP types across original and replication studies. The goal of these analyses was to see if there are any type specific carry-over effects in QMPs between original and replication. Due to the large number of tests, the focus was exploratory using visualization rather than inference.

Due to the large number of tests, the focus was exploratory using visualization rather than inference. It is the large number of tests combined with the fact that our sample size was not large enough to support that number of tests that we omitted these analyses from the main text. Furthermore, this analysis was intended not as supplementary to the main test for hypothesis 6. Furthermore, this analysis was intended as supplementary to the main test for hypothesis 6, and was pre-registered as exploratory. For completeness and transparency the analyses using beta-regressions are reported here.

### Model.

The model was the same as that used for the main test of hypothesis 6, only now QMP ratios for each of the five categories were used one at a time for each unique combination of original article and replication protocol QMP.

### Results.

Planned exploratory analyses were preformed to expand on the test for hypothesis 6. The relationships between original article QMP ratios and replication protocol QMP ratios were investigated for each QMP type separately. Figure 4 visually illustrates the relations of the revised QMP ratios from different QMP types between replication protocols and original articles. As shown in the figure, no relation was strong nor highly stable, and few were consistent between initial and revised coding protocols.

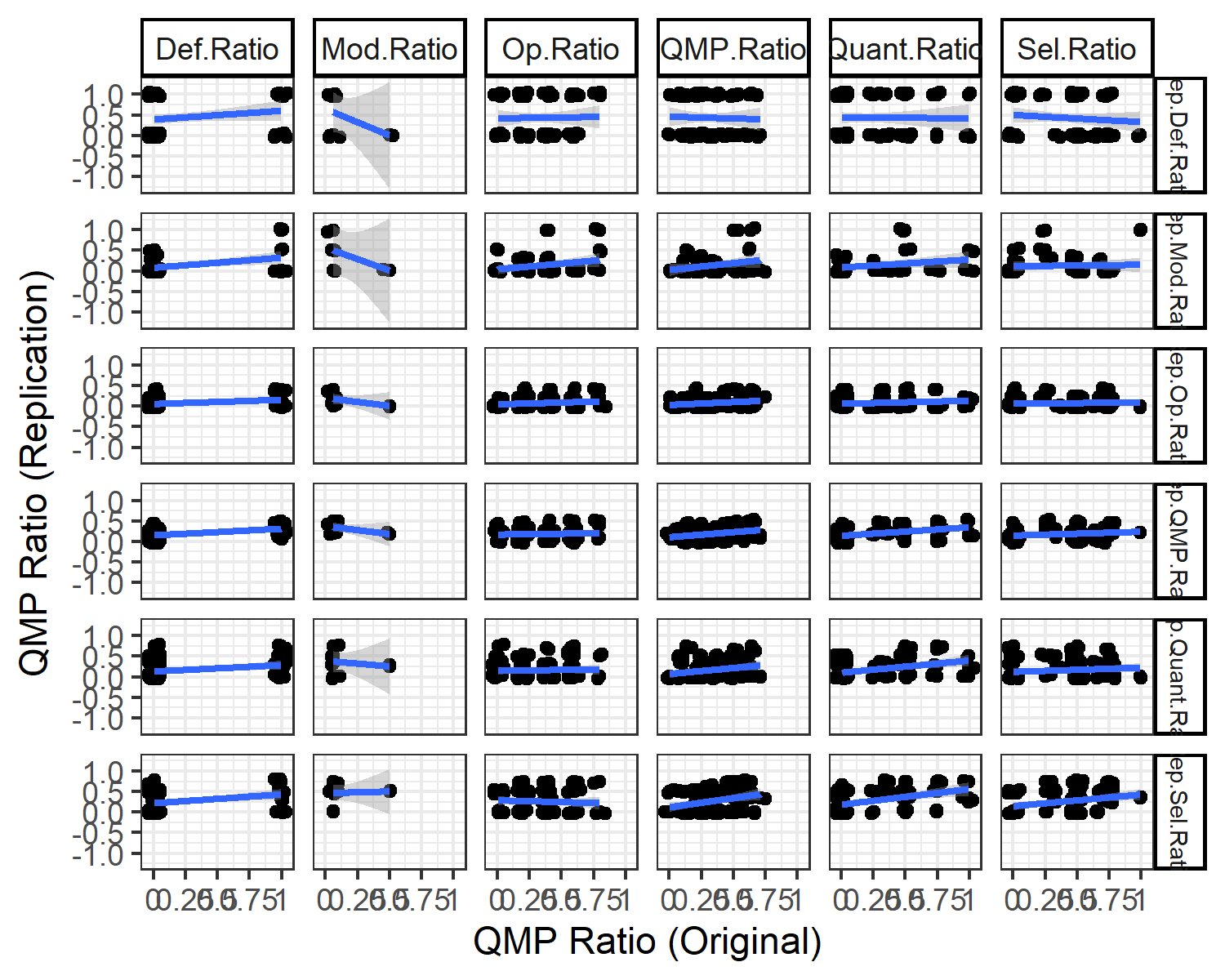


Figure 4: Scatterplot of original and replication QMP ratios per type with linear regression line. The abbreviations on the axes relate to the QMP categories described in Table 1: Def. is definition, Op. is operationalization, Sel. is selection/creation, Quant. is quantification, and Mod. is modification. The X axis facets across the QMP ratios in the original articles, and the Y axis facets across the QMP ratios in the replication protocols. Each dot in the figure relates to describes the QMP ratio for the types in that graph across an original article and its replication protocol. Note: jitter was applied to these dots to show the number of observations at points where multiple dots were present

Table 4 shows the beta-regression tests for each of the QMP category comparisons. The relations that are consistent are those between original article definition QMPs with replication selection and total QMPs, original article selection QMPs with replication total QMPs, and finally original article quantification QMPs with replication selection QMPs. All these relations were positive across both coding protocols (see Appendix D), illustrating that a greater QMP ratio of the specified type in an original article is associated with an increase in QMP ratio for the specific QMP type, or total QMP ratio in the replication protocol.

Table   
 *Beta Regression Test Results for the Association between Replication QMP Ratio and Original QMP Ratio as obtained using the Revised Protocol*

| Original |  | SD |  |  | SD |  |
| --- | --- | --- | --- | --- | --- | --- |
| QMP Type |  | Definition |  |  | Quantification |  |
| Definition | 0.46 | 0.40 | .254 | 0.60 | 0.31 | .057 |
| Operationalisation | 0.09 | 0.75 | .904 | 0.04 | 0.59 | .946 |
| Selection | -0.36 | 0.55 | .509 | 0.51 | 0.43 | .235 |
| Quantification | -0.03 | 0.55 | .960 | 1.15 | 0.43 | .007 |
| Modification | -2.58 | 3.21 | .420 | 0.36 | 2.71 | .893 |
| QMP Type |  | Operationalisation |  |  | Modification |  |
| Definition | 0.50 | 0.28 | .071 | 0.75 | 0.46 | .102 |
| Operationalisation | 0.43 | 0.54 | .419 | 0.88 | 0.75 | .240 |
| Selection | 0.08 | 0.39 | .844 | 0.06 | 0.60 | .916 |
| Quantification | 0.39 | 0.39 | .311 | 0.46 | 0.56 | .413 |
| Modification | -2.55 | 2.46 | .301 | -3.25 | 3.16 | .303 |
| QMP Type |  | Selection |  |  | Total |  |
| Definition | 0.77 | 0.33 | .021 | 1.05 | 0.23 | < .001 |
| Operationalisation | -0.52 | 0.62 | .402 | -0.13 | 0.50 | .796 |
| Selection | 1.23 | 0.45 | .006 | 0.75 | 0.36 | .038 |
| Quantification | 1.52 | 0.46 | < .001 | 1.14 | 0.34 | < .001 |
| Modification | 0.88 | 2.46 | .719 | -1.78 | 1.28 | .164 |

*Note.* Rows indicate Original QMP type, and within table indicates Replication QMP type. Sample size for comparisons involving modification were lower since not all measures were modified. Therefore, the standard errors are larger.

A possible explanation for these results could be that original articles have to provide a clear definition as well as details on the scale, in order for a replication to be able to justify selecting the measure. A replication may also be in general more dependent on the original article for information on the definition of the constructs and the choice of measurement, compared to a description of the operationalisation, quantification, and any modifications. These former are features of a measurement, which while preferably similar in a replication to the original, can and should be reported in the replication protocol even if the original article does not contain the details on these features itself. However, these are just speculations as the results showed no clear relations between different QMP types across original and replication (see Figure 4), the number of tests were many while multiple testing was not corrected for, and finally testing of causal links was not possible with this data.

# Supplementary Analyses B: Multilevel Analyses

This appendix contains the multilevel model versions of the models used to test Hypothesis 3, 5, & 6. These models were excluded from the main text for the sake of brevity, and because the sample size was not large enough to support multilevel modelling.. Guidelines suggest to have at least within the range of 10 to 30 groups with 10 to 30 cases each for multilevel models, with more groups and cases each being necessary for more complex models, such as random-slope models (Kreft, 1996; Snijders and Bosker, 1993; Hox, 1998, 2010). For Hypothesis 3 we had `r nrow(table(data\_h3\_multiple$g))` groups with on average `r round(mean(table(data\_h3\_multiple$g)),3)` cases. For Hypothesis 5 and 6 we had `r nrow(table(coded\_data\_replications$many\_labs\_version))` groups with on average `r mean(table(coded\_data\_replications$many\_labs\_version))` cases. Therefore, while we have may have had a sufficient number of cases per group for simple models, we likely lacked a sufficient number of groups to estimate multilevel models, especially more complex models. In 4 (all random slope models) of the 12 multilevel models we ran we encountered convergence issues, further illustrating that our sample size was likely inadequate for these models.

Additionally, our dependent variable (replication success) was at the group level in all of these models. In this case, it is advised in most cases to run analyses with all variables aggregated at the group level and adjust the standard errors of the coefficients using White's heteroscedasticity adjustment (Foster-Johnson & Kromrey, 2018). We therefore also show analyses with this adjustment for Hypothesis 3. Aggregation would make little sense for Hypothesis 5 and 6, as there were only four groups here, which would lead to a model with only 4 cases.

## Hypothesis 3, Nested Within Replication Sets

### Explanation of Relevance.

In this data, replication attempts at different lab locations can be seen as nested within a replication set. To see if this nesting might impact the analyses of hypothesis 3, multilevel models are included here as sensitivity checks.

### Model.

The specific models used were a multilevel logistic regression random intercept and random slope model. The dependent variable remains replication success at the replication set level, while Cronbach’s alpha was no longer averaged across locations.

### Results.

The results of these analyses can be seen in the table below. In none of these models is the relation between reliability and replication success significant (random intercept: , 95% CI , , ; random slope: , 95% CI , , ). This corroborates the findings in the article, only now using a random intercept multilevel model. However, also similarly to the results in the article and perhaps even more so for the multilevel model, the small sample size for these tests means that caution should be taken in interpreting these results.

Table   
 *Model Results from Tests of Hypothesis 3*

| Term |  | 95% CI |  |  |
| --- | --- | --- | --- | --- |
| Random-Intercept Multilevel Logistic Model |  |  |  |  |
| Intercept | -17.74 | [-36.93, 1.45] | -1.81 | .070 |
| Alpha | 1.70 | [-20.96, 24.36] | 0.15 | .883 |

*Note.* The random slope model for Cronbach’s alpha did not converge, thus its results are not shown here.

## Hypotheses 5 & 6, nested within Many Labs

### Explanation of Relevance.

The models used in the article to test Hypotheses 5 & 6 had independence of observations as an assumption, the observations here referring to a replication set. this is known to be false since the data is nested within four separate Many Labs projects.

### Models.

To test if this nested structure might influence the relations described in Hypothesis 5 & 6, random intercept, and random slope multilevel models with each Many Labs project as a group were implemented as sensitivity analyses. As of writing this article the *betareg* package does not allow for estimation of multilevel models. As an approximation, Gaussian random-intercept multilevel models were implemented using the *lmer* function from the *lme4* package (Bates et al., 2015). The outcome QMP ratio variable was transformed using the logit link function (the default link function in the *betareg* package) to better fit the Gaussian model.

### Results.

For Hypothesis 5, the sensitivity analyses using a random-intercept multilevel regression found a similar significant association as that found by the main test (, 95% CI , , ). This result is in line with hypothesis 5. The random slope model did not converge.

For Hypothesis 6, the sensitivity analyses using a random intercept logistic multilevel model resulted in non-significant results, which is different from the results found with the main test (revised protocol: , 95% CI , , ). This result does not corroborate Hypothesis 6. The random slope model did not converge.

**References**

Foster-Johnson, L., & Kromrey, J. D. (2018). Predicting group-level outcome variables: An empirical comparison of analysis strategies. \*Behavior Research Methods\*, \*50(6)\*, 2461–2479. <https://doi.org/10.3758/s13428-018-1025-8>

Hox, J. (1998). Multilevel Modeling: When and Why. In: Balderjahn, I., Mathar, R., Schader, M. (eds) Classification, Data Analysis, and Data Highways. Studies in Classification, Data Analysis, and Knowledge Organization. \*Springer, Berlin, Heidelberg\*. <https://doi.org/10.1007/978-3-642-72087-1\_17>

Hox, J., Moerbeek, M., & van de Schoot, R. (2010). Multilevel Analysis: Techniques and Applications, Second Edition (2nd ed.). \*Routledge\*. <https://doi.org/10.4324/9780203852279>

Kreft, I. G. (1996). Are multilevel techniques necessary? An overview, including simulation studies. \*Unpublished manuscript, California State University, Los Angeles\*.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard Errors and Sample Sizes for Two-Level Research. \*Journal of Educational Statistics\*, \*18(3)\*, 237-259. <https://doi.org/10.3102/10769986018003237>

# Supplementary Analyses C: Omega Coefficient Analyses

This appendix contains all results based on calculated reliabilities using the omega coefficient as the coefficient of choice instead of the alpha coefficient.

### Explanation of Relevance.

Omega was included since it has been argued to be a more informative measure of reliability than alpha, while also providing validity evidence for the scale (Crutzen & Peters, 2017; Deng & Chan, 2017). We focused our attention on the Cronbach’s Alpha in the main text, because Cronbach’s Alpha is used more commonly in research allowing us to make comparisons between calculated and reported reliabilities. Furthermore, we were able to calculate the standard error of alpha to be used in the Reliability Generalization meta-analysis. Omega was calculated using the omega function in the *psych* R package [\*v`r getNamespaceVersion("psych")[[1]]`\*; @R-psych]. Default arguments were used in the function except the nfactors argument, which was set to 1.

### Models.

To separate between and within variance in reliability, a multilevel random intercept model was specified with replication set as the grouping variable, which was the same model used for testing. The difference was that the omega value of a replication was used as dependent variable in this instance, using the *lmer* function from the *lme4* package [\*v`r getNamespaceVersion("lme4")[[1]]`\*; @R-lme4].

The ICC was calculated similarly based on the omega as was done for Cronbach’s alpha in the article. NOALINEABREAK The analysis for H3 related the calculated average reliability coefficients for each replication set to whether the replication was successful when judged at meta-significance level of . NOALINEABREAKA multilevel random-intercept, and a random-slope equivalent model to the model above were also run for Omega, similarly to what was done for Cronbach’s alpha in [Supplementary Analyses B]( Supplementary\_multilevel\_analyses.Rmd).

### Results & Comparison Interpretation.

For omega the mean (values calculated in the population across replication sets) was 0.789 with a standard deviation of 0.163.

When looking at the between and within group variance in omega coefficients, compared to alpha an even larger relative degree of the variance was between group rather than within group when compared (between-group variance 0.004, within-group variance 0.037, ICC = 0.913). This result thus provides evidence in a similar direction to the Cronbach’s Alpha results, meaning that most variance in reliability coefficients is observed between studies rather than within studies.

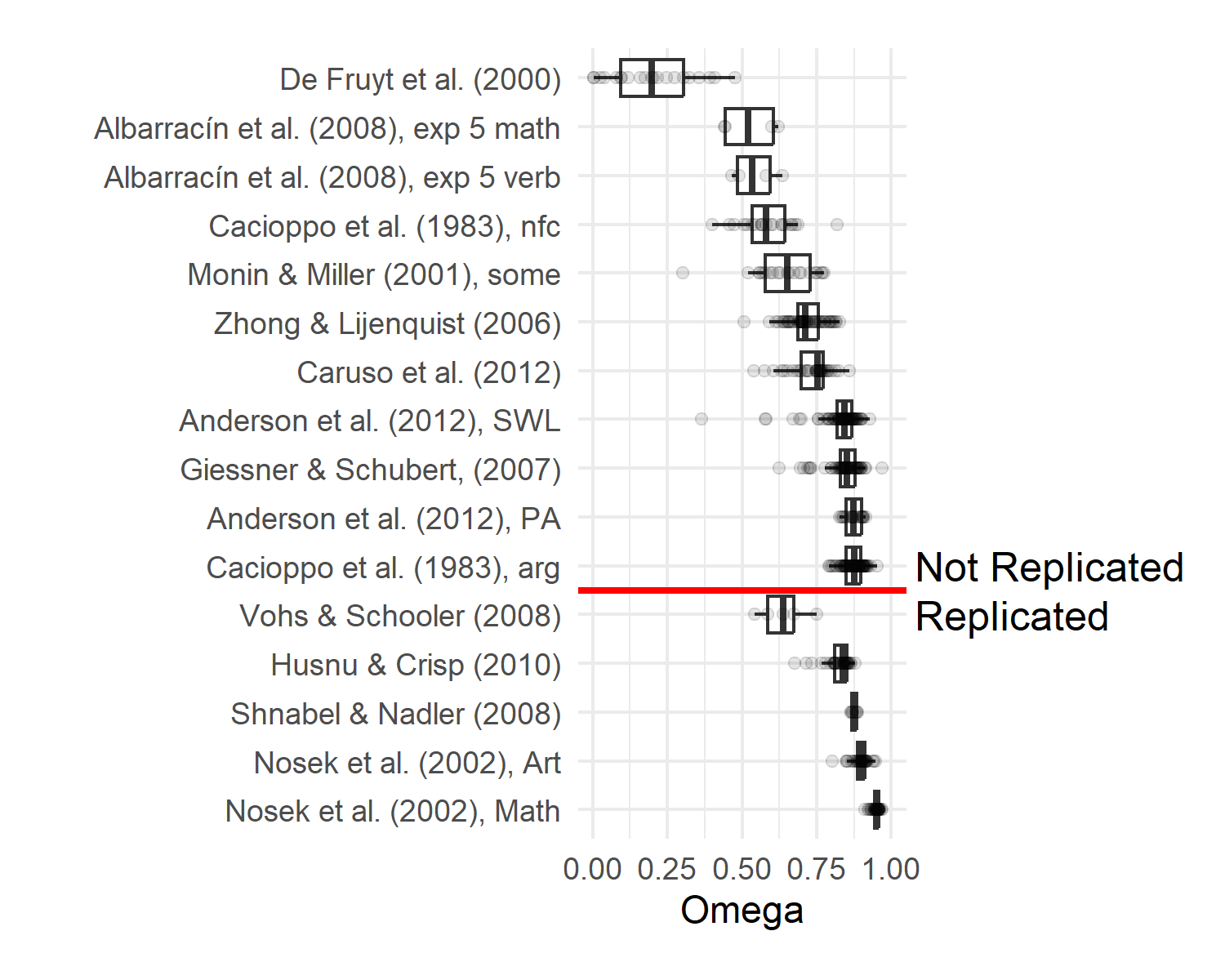
The calculated Omega coefficient did not significantly relate to replication success in the main logistic regression model. Similarly, non-significance was also observed in the multilevel random-intercept and random-slope models (random intercept: , 95% CI , , ; random slope: , 95% CI , , ). The random slope model did not converge.

Table   
 *Model Results from Tests of Hypothesis 3 based on the Omega Coefficient*

| Predictor |  | 95% CI |  |  |
| --- | --- | --- | --- | --- |
| Logistic Regression Model |  |  |  |  |
| Intercept | -7.91 | [-21.15, -0.74] | -1.61 | .107 |
| Omega | 9.11 | [0.07, 24.76] | 1.53 | .126 |
| Random-Intercept Multilevel Logistic Model |  |  |  |  |
| Intercept | -21.06 | [-77.85, 35.73] | -0.73 | .467 |
| Omega | 5.91 | [-61.58, 73.39] | 0.17 | .864 |
| Random-Slope Multilevel Logistic Model |  |  |  |  |
| Intercept | -17.03 | [-65.84, 31.78] | -0.68 | .494 |
| Omega | 0.90 | [-57.21, 59.02] | 0.03 | .976 |

Thus, no evidence in favor of Hypothesis 3 was found. However, caution in interpreting the test results is needed. The number of measures for which the omega coefficient could be and was calculated was quite low (sixteen).

The figure below shows additional information on the distributions of the calculated omega scores and their relation to replication success. It mostly resembles Figure 1 from the article, with some studies being omitted as the omega coefficient function did not converge or run properly on those data. Furthermore, information regarding heterogeneity was not present since the standard error of omega was not known, and comparisons between reported coefficients were omitted, since no article reported any omega coefficients.



# Supplementary Analyses D: Initial Coding Protocol QMP Analyses

This appendix contains the QMP analyses that in the article and appendices were based on the revised QMP, but then performed based on the QMPs obtained using the initial coding protocol.

### Explanation.

The data was initially coded using to the preregistered coding protocol. After the coding was completed using the preregistered coding protocol, it was determined that some of the initial ratings were possibly too stringent. As a result, some of the items in the protocol were reformulated, so that the conclusions from the resulting analyses were based on a more lenient judgement of QMPs. The coded data using the initial protocol was still separately available, and because this protocol was the one that was pre-registered the results for those analyses are shown here.

### Models.

To test hypothesis 4, the total QMP ratios between replication and original research was compared using a beta-regression where total QMP ratio was predicted by whether the QMP ratio came from an original article, or a replication protocol.

For hypothesis 5, a logistic regression model was used to test the association between the QMP ratio in replication protocols and replication success. In this appendix we run the model based on the QMP ratios obtained using the initial protocol. NOALINEABREAKThe relation between QMP ratios in the replication protocols and corresponding original articles were investigated using a beta regression model. NOALINEABREAK[Supplementary Analyses B](Supplementary\_multilevel\_analyses.Rmd)already contained the results of the multilevel sensitivity models for Hypothesis 5 & 6 based on the QMP ratios obtained using the revised coding protocol. In this appendix we show the result for the same models based on the QMP ratios obtained using the initial protocol.

### Results & Comparison Interpretation.

The figure below displays the distribution of observed QMP ratios across original articles and replication protocols based on the initial coding protocol.

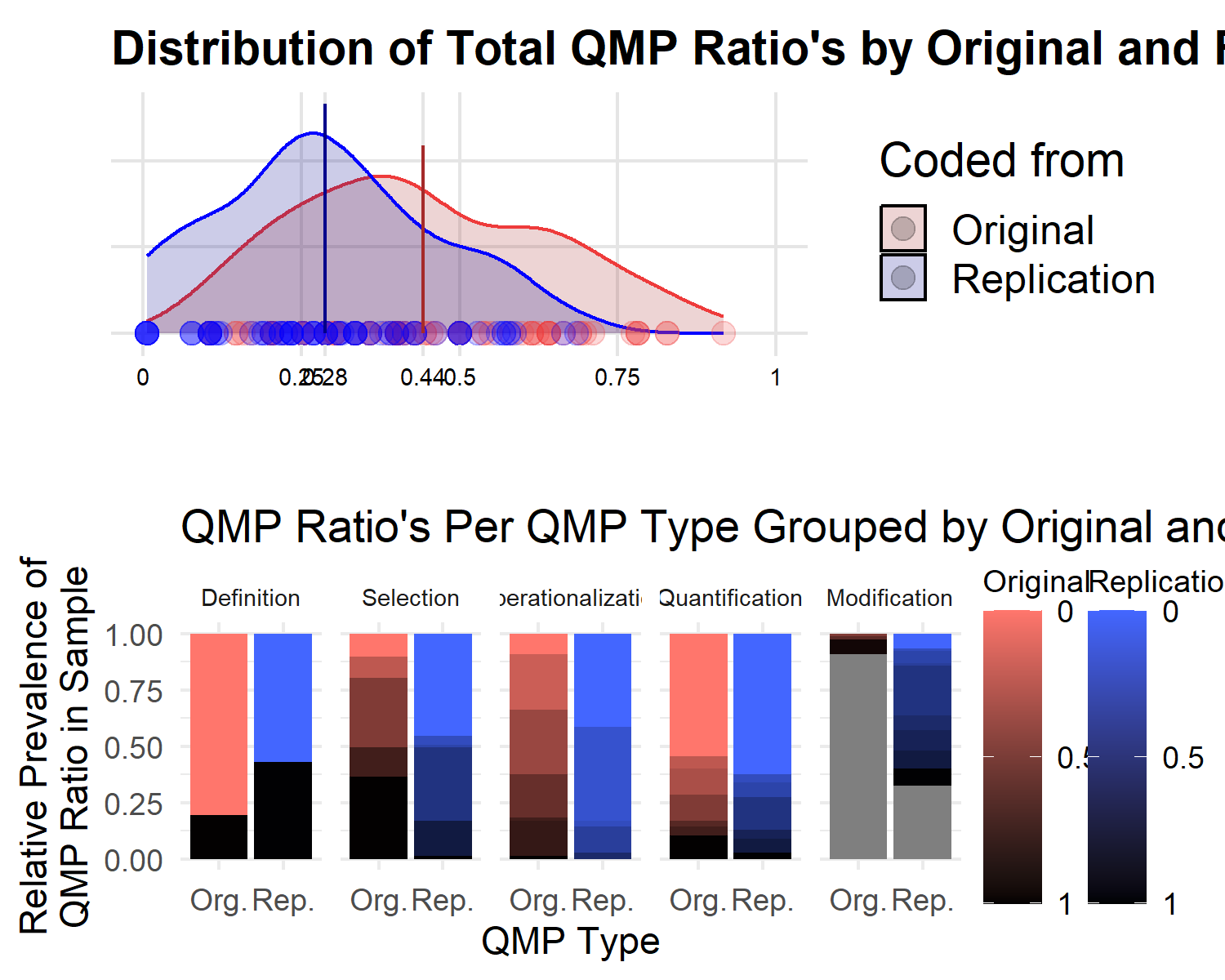


Figure 5: QMP ratio counts for each QMP Type and QMP total ratio distribution grouped by whether the QMP ratio was obtained from an original article or a replication protocol. The top graph shows the distributions of total QMP ratios for both replication protocols and original articles, with the line indicating the mean QMP ratio. The specific observed values are indicated along the bottom row with dots. The bottom row shows for each QMP type the proportions of each QMP ratio obtained, darker colors represent proportionally more QMPS, grey means modification did not occur for that measure..

The bottom panel of Figure 5 already shows that original articles contained a significantly larger proportion of QMPs than replication protocols for the measures of the same effects. Beta-regression was used to test the difference in QMP ratio based on the initial protocol coding, between original articles and replication protocols. The results of this test indicated that this difference was significant (, 95% CI , , ). This result is similar to the result found using the revised coding protocol in line with hypothesis 4.

Contrary to the results based on the revised coding protocol, a decrease in QMP ratio based on the initial coding protocol did not significantly relate to replication success (, , 95% , , ). Thus, not supporting hypothesis 5.

Similarly to the previous test, the relation was not significant for the QMP ratios obtained with the initial protocol (, 95% CI , , ). Thus, not supporting Hypothesis 6. Figure 6 displays this relationship visually.

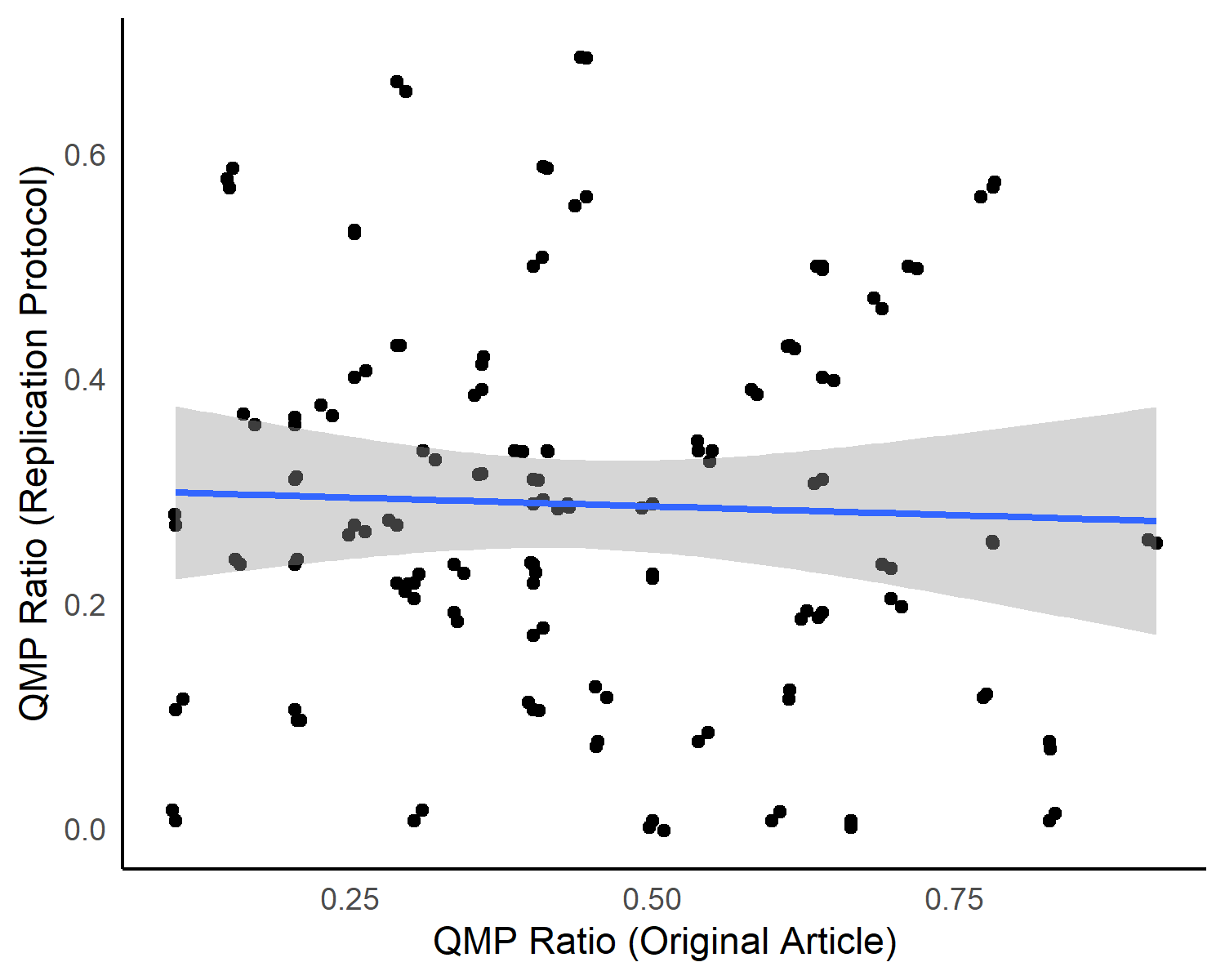


Figure 6: Scatterplot of original and replication total QMP ratios with linear regression line. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots to show the number of observations at points where multiple dots were present. NOALINEABREAK

Table   
 *Beta Regression Test Results for the Association between Replication QMP Ratio and Original QMP Ratio as obtained using the Initial Protocol*

| Original |  | SD |  |  | SD |  |
| --- | --- | --- | --- | --- | --- | --- |
| QMP Type |  | Definition |  |  | Quantification |  |
| Definition | 0.46 | 0.40 | .254 | 0.47 | 0.35 | .174 |
| Operationalisation | -0.32 | 0.63 | .612 | -0.16 | 0.54 | .767 |
| Selection | -0.40 | 0.48 | .403 | 0.14 | 0.40 | .736 |
| Quantification | 0.17 | 0.47 | .709 | 0.60 | 0.40 | .134 |
| Modification | -0.88 | 3.09 | .775 | -4.65 | 2.72 | .087 |
| QMP Type |  | Operationalisation |  |  | Modification |  |
| Definition | 0.53 | 0.28 | .064 | -0.71 | 0.73 | .331 |
| Operationalisation | -0.02 | 0.46 | .966 | 0.78 | 1.47 | .597 |
| Selection | -1.07 | 0.34 | .002 | -1.49 | 0.66 | .024 |
| Quantification | 0.10 | 0.34 | .758 | 2.34 | 0.94 | .013 |
| Modification | -0.68 | 2.05 | .741 | 0.44 | 1.56 | .778 |
| QMP Type |  | Selection |  |  | Total |  |
| Definition | 0.90 | 0.35 | .010 | 0.72 | 0.25 | .003 |
| Operationalisation | -0.54 | 0.54 | .320 | -0.66 | 0.41 | .109 |
| Selection | 0.11 | 0.41 | .791 | -0.89 | 0.30 | .003 |
| Quantification | 1.09 | 0.41 | .007 | 0.53 | 0.30 | .076 |
| Modification | -4.18 | 2.77 | .132 | -1.36 | 1.74 | .436 |

*Note.* Rows indicate Original QMP type, and within table indicates Replication QMP type. Sample size for comparisons involving modification were lower since not all measures were modified. Therefore the standard errors are larger.

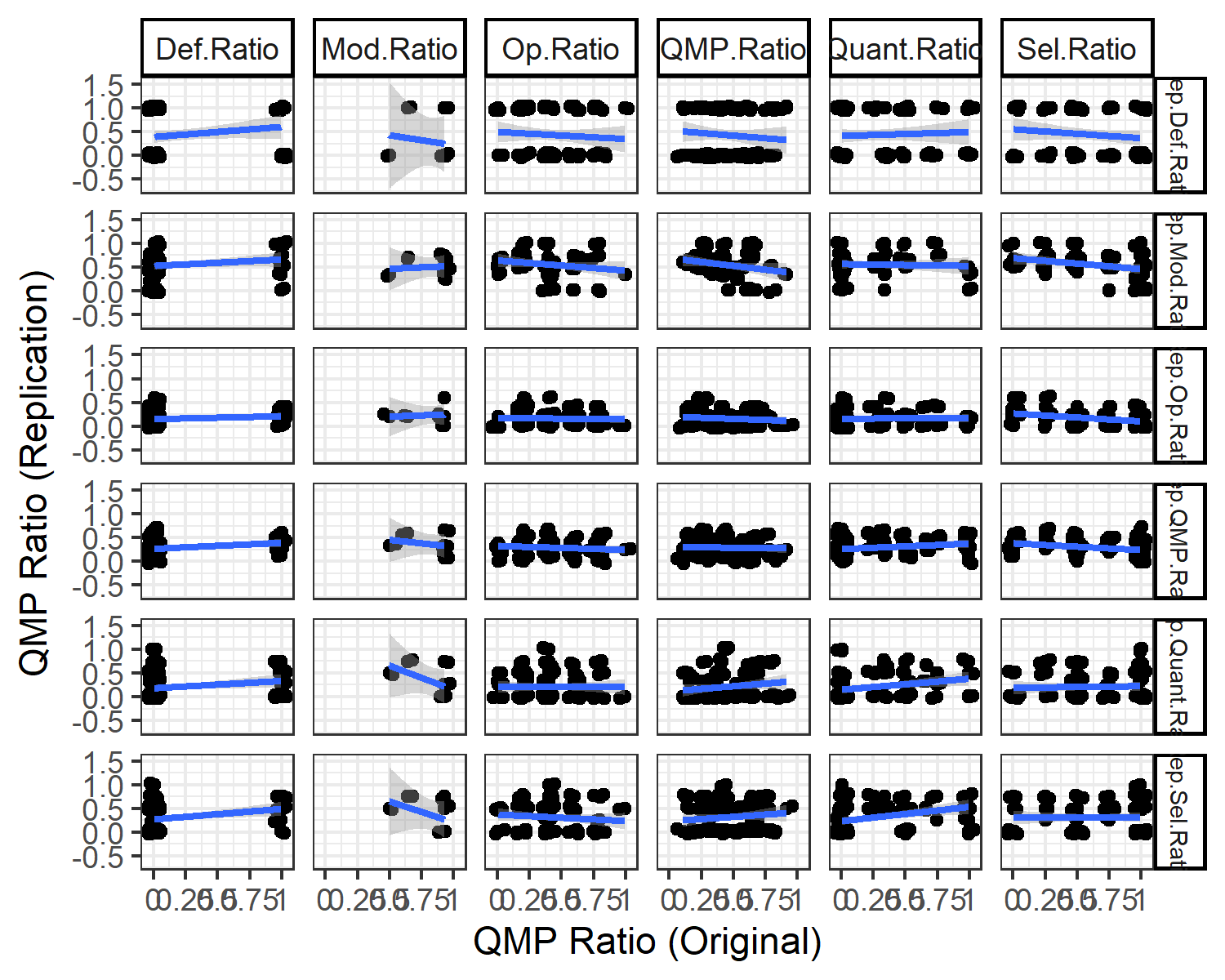


Figure 7: Scatterplot of original and replication total QMP ratios obtained using the intial coding protocol with linear regression line. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots to show the number of observations at points where multiple dots were present

Now we look looking at the multilevel sensitivity analyses, only now using the QMP ratios based on the initial coding protocol. NOALINEABREAKContrary to the results of the multilevel sensitivity model testing hypothesis 5 using QMP ratios obtained with the revised coding protocol, the relationship in the random intercept and random slope was no longer significant for QMP ratios obtained with the initial coding protocol (random intercept: `r random\_intercept\_h5\_REV$full\_result$hypothesis\_supportTRUE`; random slope: `r random\_slope\_h5\_REV$full\_result$hypothesis\_supportTRUE`).

For Hypothesis 6, similarly to the results of the multilevel sensitivity model testing Hypothesis 5 using QMP ratios obtained with the revised coding protocol, the relationship in both the random intercept and random slope models was not significant for QMP ratios obtained with the initial coding protocol (random intercept: `r random\_intercept\_h6\_REV$full\_result$QMP`; random slope: `r random\_slope\_h6\_REV$full\_result$QMP`).

1. Initially the original articles contained 3 more measures than the replication protocols. This difference was due to the way that the moral foundations questionnaire was framed in the original articles compared to in the original article. In the original article it was framed as measuring five different moral foundations, while in the replication protocol the measure assessed the two overarching categories that were used to test the main effect in both the original and replication research. The measurement information reported was comparable across all five categories, and thus it was deemed that the measurement could be reduced to reflect two overarching categories facilitate easier comparison between measurement in original and replication. [↑](#footnote-ref-1)