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. Goos, C.: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Bakker, M.: Supervision, Writing - Review & Editing; Wicherts, J.M.: Supervision, Writing - Review & Editing; Nuijten, M.B.: 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: [casgoos99@gmail.com](mailto:casgoos99@gmail.com)

Abstract

Psychological science has been facing threats to the credibility of its findings, particularly due to the lack of replicability of those findings. Recently, attention has increased towards the impact that challenges in psychological measurement have on replicability. In this article, we assessed the reliability and measurement reporting within the Many Labs replications and related original studies. We then empirically investigated the associations of replicability with reliability and measurement reporting. The results indicated that not all measures were sufficiently reliable across contexts. Additionally, reliability and validity evidence was rarely reported. Finally, incompleteness in measurement reporting was associated with lower replicability of the findings tested with that measure. These findings indicate a worrying lack of reported validating evidence for measurement in the literature, as well as the potential harm of incomplete measurement reporting for establishing credible findings. The article ends with suggestions on improving measurement validation and reporting practices. Finally, we suggest that researchers should evaluate the reported measurement information when deciding what studies to replicate.

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

*Word count:* 164

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

# Introduction

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 affected by problems in psychological measurement, and in turn, how those problems may affect successful replication.

### Credibility & Replicability.

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

In practice, effects deviate substantially between original and replication research. In 2015, the Replication Project: Psychology (RPP) attempted replications of 100 research findings in psychology, they observed a substantial discrepancy between the replication findings as compared to the original findings (Open Science Collaboration, 2015). Similarly, the Many Labs projects replicated the effects reported in original articles across different labs in various locations around the world (Camerer et al., 2016; Camerer et al., 2018; Ebersole et al., 2016, 2020; Klein et al., 2022, 2014, 2018). In these replications it was observed that effects were on average smaller in replications compared to original research. Furthermore, there was variation within the estimates of the effect 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). Simultaneously researchers sought to explain this discrepancy. The factors that have been proposed include, questionable research practices (Cumming, 2014; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016), questionable reporting practices (Bakker & Wicherts, 2011; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016), and biased publication decisions (Bakker, Van Dijk, & Wicherts, 2012; Giner-Sorolla, 2012; Sterling, 1959). A factor that has gained more traction recently is the impact of measurement-related challenges on replicability.

### Credibility & Measurement.

Measurement in psychological science is characterized in large part by one overarching issue: psychological constructs, such as affective state or intelligence, can usually not be measured directly (Flake, Pek, & Hehman, 2017). As a result, psychologists face two primary challenges in the use of measurement: First of all, because the measurement is indirect, it also comes with some degree of random error surrounding its assessment. Secondly, the responses on the measure have to relate to the psychological construct(s) of interest, but as the construct is not directly observable, this is difficult to assess.

The first issue relates to the concept of score 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.

The second issue relates to the concept of validity, and specifically construct validity. Validity in measurement refers to the overall extent to which a measure measures what it is supposed to (Borsboom, Mellenbergh, & Van Heerden, 2004), while construct validity refers to the substantial relation between the scores on a measure and the related psychological construct. A researcher cannot assume to have found a credible psychological effect, if the measures do not relate to a psychological construct (Cook, Campbell, & Shadish, 2002).

### Replicability & Measurement.

Whereas both reliable and valid measurement, and replicability are important elements that independently contribute to the overall credibility of a finding, there may also be a relation between the two. Indeed, recent studies have illustrated that both the reliability of the measure (Stanley & Spence, 2014) and the reporting of the measure (Flake & Fried, 2020; Flake et al., 2017; Shaw, Cloos, Luong, Elbaz, & Flake, 2020) can impact the chance of successfully replicating a psychological finding. This study aims to investigate the influence that score reliability and measurement reporting have on replicability, using data and reports from the large scale Many Labs replication projects.

### Many Labs Sample Description.

In the Many Labs projects each effect was directly replicated across multiple lab locations over the world. As a result the data obtained from these measures allows us to say something about the variability in these measures across different contexts. Another important aspect of the Many Labs projects is that the measurement was preregistered and documented in structured protocols that were used to instruct all the participating labs. Therefore, we believe the way measurement was used in these projects represent a high standard within the field. Any problems we encounter in this sample thus means that those problems are likely equally if not more present in other psychological research.

## Existing Research

### Reliability.

While reliability is generally considered a measure specific feature, this is false. The reliability of a measure is context dependent (Cho & Kim, 2015; Pauly, Umlauft, & Ünlü, 2018), and as a result so are the scores obtained with that measure. The contextual variations in the measurement scores could explain the discrepancy between effects in original and replication research, as well as within replications of the same original effect. 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).

However, even when reliability is held equal across studies, reliability can still cause estimates of the same effect to vary significantly. 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). This implies that studies using unreliable measures have a lower probability to be successfully replicated, even if the underlying effect is true and 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 conjugate term 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 measure and procedure of any sourced measures (Flake & Fried, 2020).

Six key questions that an article should answer to avoid QMPs were identified in Flake and Fried (2020): 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, Davidson, Wong, and Pek (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. Replication researchers need to know how the measurement was conducted originally, in order to reconstruct it. If the measure is not reconstructed exactly, original and replication may be measuring different construct(s). In that case, different effect size estimates in original compared to replication research may be because the replication effect is about a different construct, and not because the original effect was not credible.

### Research Contribution.

Our study investigated the influence of both measurement reliability and measurement reporting on replicability, using data and reports from the large scale Many Labs replication projects. Firstly, this study expands on the research by Stanley and Spence (2014). Our goal was to see whether reliability was related to replication success on empirical replication data. In practical terms, we investigated whether or not replicated studies differed in their reliability scores from non-replicated studies. In their study Stanley and Spence (2014) assumed reliability to be constant across studies, which we know to be an oversimplification. Therefore, to add further context to relation between reliability and replication, this study also investigated the variation in reliability as observed across separate measurement occasions.

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 contexts. It is crucial for evaluating replications that differences between replication and original are understood and accounted for.

* RQ1a. What is the degree of score reliability in replications of psychological research?
* RQ1b. What is the degree of score reliability in original psychological research?

The Many Labs studies were intended as direct replications. In direct replications, any deviations from the original research are believed to be irrelevant for testing the same effect as the original study (Nosek et al., 2022). In that case, the reliability of measurements in replications should also not deviate systematically from the original research.

* H1. No difference is present in the degree of score reliability between replication research and original psychological research.

Reliability has the potential to vary not only between replication and original research, but also within replications of the same effect (Vacha-Haase, 1998).

* RQ2. To what degree do reliability estimates differ within a replication set of the same original study?

Heterogeneity could be a major contributor to this variation. However for effect sizes there is little empirical evidence of widespread heterogeneity among the Many Labs replications (Klein et al., 2018; Olsson-Collentine, Wicherts, & Assen, 2020). The same is expected to hold true for the heterogeneity in reliability coefficients among the Many Labs replications.

* H2. There is no significant variation in the reliability estimates of replications of the same original study.

In order to see if reliability has an impact on replication attempts, we investigated whether or not successfully replicated studies differ in the reliability of their measures as compared to non-replicated studies.

* RQ3. What is the association between replication study score reliability and replication outcome?

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.

* H3. Score reliability in replication research is positively associated with successful replication of an original finding.

### Measurement Reporting.

First of all, this study aims to conceptually replicate the findings on measurement reporting of Flake et al. (2022).

* RQ4a. To what degree are QMPs present in replications of psychological research?
* RQ4b. To what degree are QMPs present in original psychological research?

Flake et al. (2022) found that QMPs were overall more common in the RPP replications as compared to the original research. We did not hypothesize this to be the case in the Many Labs sample, because of the use of structured protocols documenting the measurement.

* H4. QMPs are expected to be more frequent in original psychological research than in replication research.

Differences in measurement resulting from QMPs have been suspected to cause replication and original effects to deviate (Flake et al., 2022). Investigating the link between QMPs and non-replicability should put this suspicion to the test.

* RQ5. What is the association between QMPs in replication studies and replication outcome?

A reduction in QMPs corresponds to greater transparency in reporting of measurement. Research on other transparency related practices have shown that these are associated with more robust estimates of effects (Protzko et al., 2020; Wicherts, Bakker, & Molenaar, 2011). We expect a similar effect for QMPs.

* H5. QMPs in replication research are negatively associated with successful replication of an original finding.

One of the reasons measurement in replication may deviate is because original articles did not report all the necessary information needed to mimic the original measurement. In this way a lack of information may carry over from the original article to the replication.

* RQ6. What is the association between QMPs in original research and QMPs in replications of psychological research?

Previous research by Flake et al. (2022) has shown tangential evidence that QMPs in original articles may cause issues for recreating measurements for replication research. Additionally, Shaw et al. (2020) found a similar spill-over effect for validity. Based on these findings, QMPs in original research were expected to be associated with QMPs in replication research.

* H6. Total number of QMPs in original psychological research is positively related to total number of QMPs in replication research.

# Disclosures

### Preregistration.

Data collection, coding protocol, and planned analyses were all preregistered. The preregistration and supplementary materials can be found on this article’s OSF page: <https://osf.io/9r8yt/>. Any 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 based on 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.

# Method

## Sample

### Data Source.

The data used for the analyses consisted of three main sources: replication datasets, preregistered replication protocols, and original study articles. The data came from the Many Labs replication projects. In particular, data from Many Labs 1, 2, 3, & 5 was used (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018). Data from Many Labs 4 (Klein et al., 2022) was not part of the sample, as there was no publicly available preregistered replication protocol. Additionally, the replication of (Crosby, Monin, & Richardson, 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.

Measures were identified using the preregistered replication protocols. Acquiescence bias checks, manipulation checks, pilot test measures, and measures added for exploratory analyses were not included. The result was a total sample size of 77 measures used in both original and replication research[[1]](#footnote-41).

## Data Collection

The data on the preregistered replication protocols, and replication datasets from Many Labs 1, 2, 3, & 5 were all retrieved 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 of the labs that took part in one of the Many Labs studies. These datasets were accessed through their OSF page. For the analyses, the scores on the items of an identified measure were identified and extracted. When scores could not be clearly identified, any available codebooks, analysis scripts or study materials were used to identify the relevant scores.

In order to be considered suitable for 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 entirely. These criteria resulted in suitable item score data from nineteen replication sets spread across on average 35 lab locations for the analyses of hypotheses 2 & 3.

### Preregistered Replication Protocols.

The preregistered 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.

The OSF page for each Many Labs project was combed through, in order to identify the documents that contained information on the measurement practice and reported reliability. For Many Labs 2, & 3 a single replication protocol for all studies could be identified. For Many Labs 1, the Proposal\_V1.1 provided the equivalent information, while for Many Labs 5 protocols were available for each set of replications separately.

The relevant OSF page of each replication set was searched through for a file which contained protocol in its name. If available, a version of the protocol labelled as revised, post-review, peer-review, or endorsed was selected. Otherwise, the latest uploaded relevant protocol was selected. Protocols labelled as data collection, analysis protocol or anything similar were excluded[[2]](#footnote-50).

Finally, the published reports of the Many Labs projects (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018) were used to determine whether not the replication was considered successful.

### Original Articles.

The original study articles are the published articles of the original effect on which the Many Labs replication sets were based. They were identified using the citations for these articles in each replication protocol. In the end all articles could be retrieved through searches on Web of Science, Google Scholar, and PsychInfo (in that order)[[3]](#footnote-53).

## Measures

A structured coding protocol was used to extract information on the reported measurement evidence and reporting quality for both replication protocols and original studies. The coding protocol and explanation of its use can be found in the coding protocols folder of the supplementary materials.

### Measures of Reliability.

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

Cronbach’s alpha was calculated as the main score reliability index used in the analyses, as it remains the most commonly used indicator of the reliability of a measure (Flake et al., 2017). This is important for facilitating comparison between the calculated replication reliabilities and the reported original article reliabilities. Cronbach’s alpha was calculated using the alpha function from the psych R package (William Revelle, 2023). Default arguments were used for the function. Omega was pre-registered as an additional reliability index score, 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.

In order to answer questions regarding reliability, the reported reliability coefficient of a measure was extracted from both the replication protocols and the original articles when present. Both the type (Cronbach’s alpha, retest, interrater, etc.) and the value itself were extracted. Similarly, we coded the presence of any validity evidence, such as a factor analysis, alongside the measure. The aim here was to gauge the presence of factor analyses within the articles and protocols, for comparison with Shaw et al. (2020).

The tests for hypotheses 4, 5, & 6 were all based on the QMPs coded for both original articles and replication protocols. The QMPs we coded were in part based on Flake et al. (2022). Additional items were added to more fully reflect the broader questions regarding transparent measurement reporting 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). An overview of the QMP categories and 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 considered 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 particular measure (e.g. reporting results from a factor analysis for single item measures). For the analyses a ratio was calculated based on the responses to the QMP questions. This ratio was based on the number of true responses divided by the number of responses coded to be applicable. The ratio was calculated for each QMP type and across all QMPs together.

After the initial coding was completed, we made a small revision to the preregistered coding protocol. The protocol was changed because some items were considered too stringent in 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. The analyses, tables and figures presented in this article are all based on the revised coding protocol, the equivalent results based on QMPs obtained with the initial protocol can be found in [Appendix D](AppendixScripts/Appendix_initial_QMP_analyses.Rmd).

### Measure of Replication Success.

Replication success was determined based on the significance of the meta-analytic effect of the replication set as reported in its respective Many Labs report. An effect was considered 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 to the authors’ knowledge have been proposed in the literature, but not yet empirically tested on real data. In this exploratory context, false negatives were considered more harmful than false positives.

# Results

## Descriptives

The number of replication sets, measures, and measures per study that were extracted, as well as the ratio of successful replications can be seen in Table 2, 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 succesfully. 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.

The number of participants in the replication studies, and the associated articles are displayed in Figure 1. As might be expected, the Many Labs replications always had larger sample sizes than the original articles. Furthermore, the figure shows that the replications are nested in Many Labs projects, and that within some replications multiple measures were used.

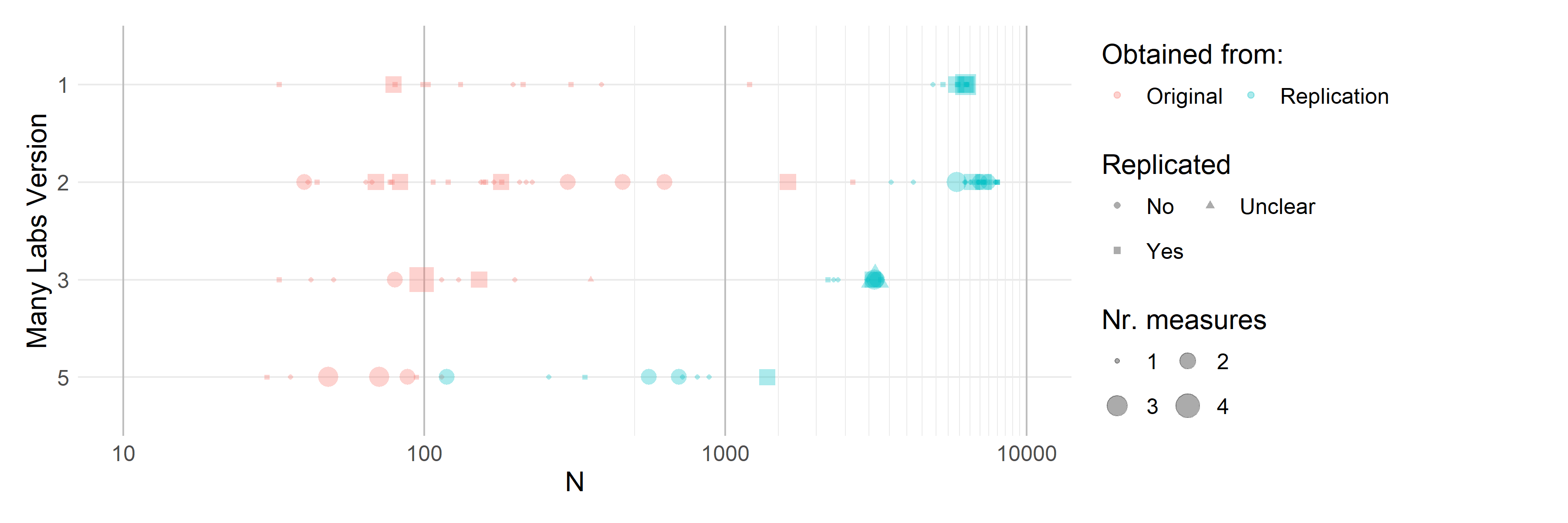


Figure 1: This figure displays the sample size for the included original articles and total sample size for multiple replications. The color of each point indicates whether it represents an original article or a replication. Each point’s shape relates to whether the hypothesis was supported or not, while the size indicates the number of measures used within that article or replication. The figure is log10-scaled, each light grey vertical line represents a step of 500 in sample size. Two original articles are not represented in this figure due to them containing no reported sample size.

## Measurement Reliability

If data from a multiple item scale could be accessed, the Cronbach’s Alpha of that scale was calculated 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, López-Nicolás, Blázquez-Rincón, & Sánchez-Meca, 2024; Vacha-Haase, 1998). This enabled us to quantify the degree of true variation (or heterogeneity) in reliability coefficients across lab locations.

A meta-analysis requires an estimate of the standard error. For Cronbach’s alpha, formulas 2 & 3 from Duhachek and Lacobucci (2004) were used to calculate its standard error. 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 (Viechtbauer, 2010). Defaults settings were used. No correction for bias was implemented, 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 2 displays the distributions of the calculated Cronbach’s Alpha scores from each lab for each measure, separated by successful and unsuccessful replication.

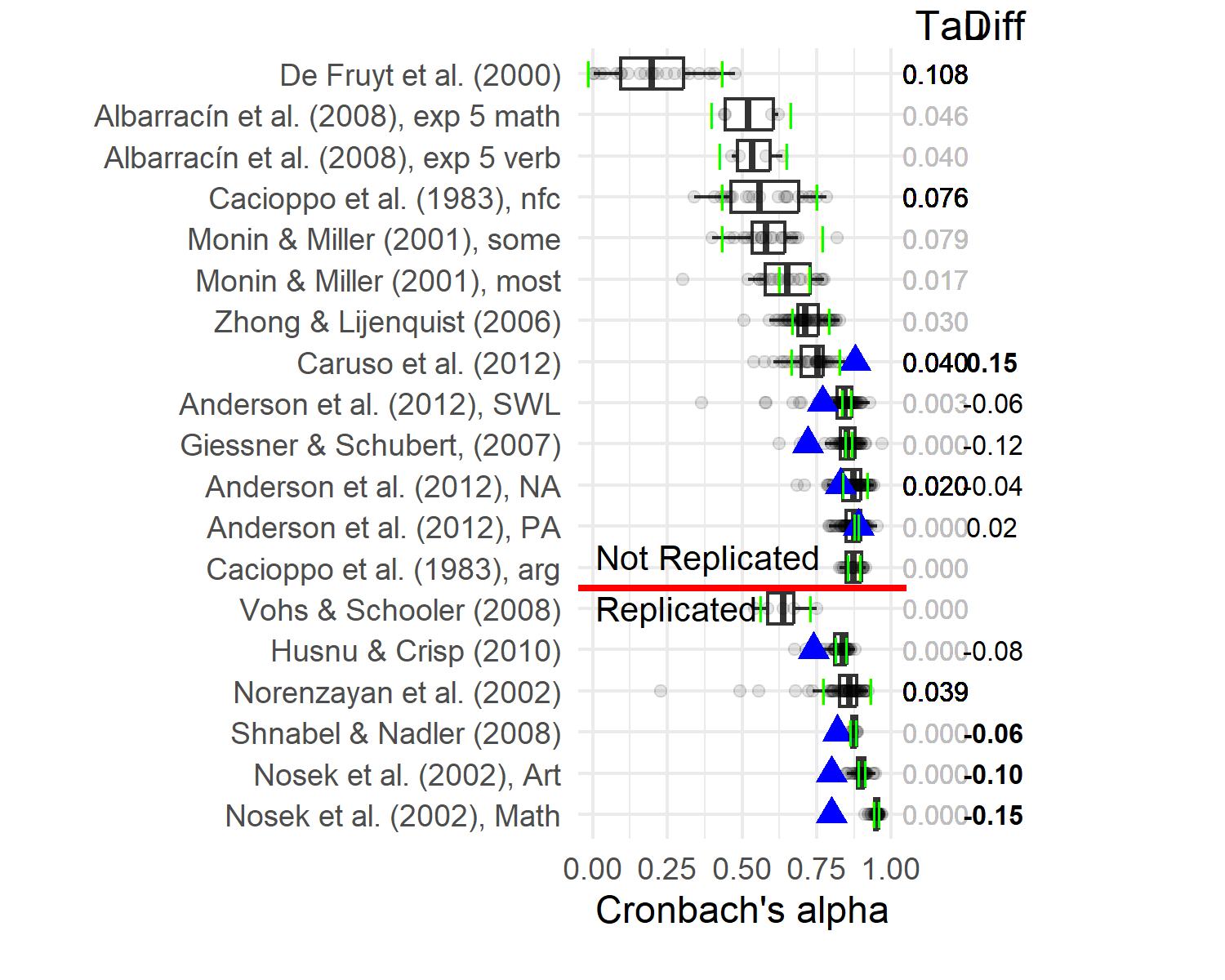


Figure 2: Distributions of calculated Cronbach’s alpha coefficients (> 0) calculated for the responses on a measure at each lab location, across the eighteen distinct measures for which suitable raw data was available to calculate Cronbach’s alpha coefficients from. The green lines indicate the meta-analytic prediction interval lower and upper bound. The blue triangles indicate the reported alpha coefficient for that measure from the original article, when available. 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 the measures varies highly. Most of the measures near the bottom show average reliability scores of at least .80, corresponding to an adequate reliability for general research purposes (Nunnally & Bernstein, 1994), with minimal variation between labs. However, other measures show not only considerably lower average reliability scores, but also greater variation.

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

### H1.

There were not enough reported reliabilities test whether or not the value of the reported Cronbach’s Alpha was different for replications compared to original articles, using the pre-regsitered Mann-Whitney U test. However, the difference between the reported reliability in the original article and the calculated average reliability in replications can be seen in the Diff column in Figure 2. 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.

### H2.

We used the RG meta-analyses to test if there was heterogeneity in these estimates, to get an indication of the true variability in the reliability scores. This analysis was not preregistered, because we were not aware of it at the time of preregistering.

Figure 2 displays the tau score for each measure’s Cronbach’s Alpha, which indicates the true differences in reliability scores between labs. For some measures the tau estimate was approximately 0. But this was certainly not the case for all measures. For instance, for the top measure the Tau estimate was .108. This means that the standard deviation of the true reliability of the measure is equal to .108 points of Cronbach’s Alpha. Figure 2 shows that for 5 out of the 19 measures, the Q-test for heterogeneity indicated that the true variation in reliability estimates was significant.

These results indicate that for several measures in our sample, the reliability of the measures is heterogeneous, thus showing evidence in favor of rejecting H2. However, most measures did not show evidence of significant reliability heterogeneity.

### H3.

Figure 2 additionally indicates whether or not the measures were part of a successful or failed replication. A logistic regression model was used to test whether there was a significant difference in the average calculated reliability coefficient value based on whether or not the associated replication was successful or not based on the meta-significance of the effect.

Contradictory to hypothesis 3, the logistic model indicated that Cronbach’s alpha did not significantly relate to replication success in the main logistic regression model (Cronbach’s alpha: , , 95% CI , , ). However, it should be noted that the reliability coefficient could be calculated for only a small sample number of 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 3 depicts the flow of measures in relation to reliability reporting. First of all, 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 twelve in the original articles reported a reliability coefficient. The most commonly reported reliability coefficient was Cronbach’s alpha, which accounted for three out of four reported reliabilities in the replication protocols, and eleven out of thirteen in the original articles.

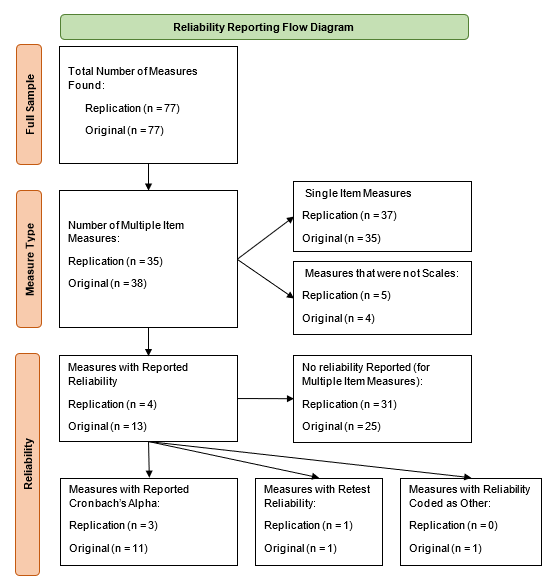


Figure 3: Reliability reporting flow diagram. Figure shows the number of measures as reported in both the replication protocols and original article, which meet the criterion in the box within the diagram and those criteria before it.

Reliability was more commonly reported in original as compared to replication research ((1) = 4.09, *p* = .043). Validity evidence was similarly reported infrequently. The number of original articles that reported validity evidence from a factor analysis was 8, for replications this number was 3.

### H4.

The bottom of Figure 4 displays the distributions of the total QMP ratio for both original articles and replication protocols. Here we can see that the average QMP ratio in replication protocols is smaller (0.18) than in original articles (0.31). The top of Figure 4 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.

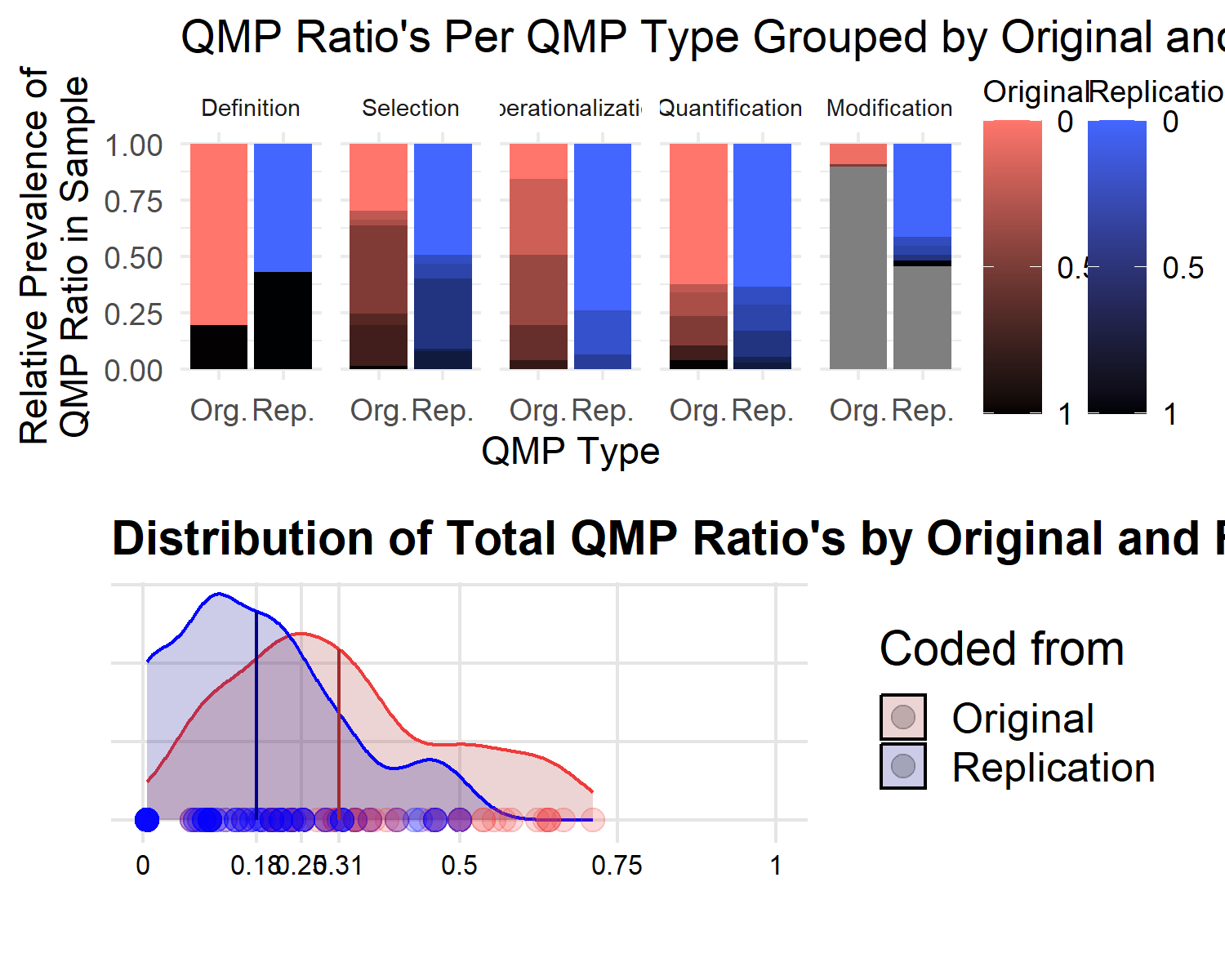


Figure 4: QMP ratio counts for each QMP Type and QMP total ratio distribution grouped by whether the QMP ratio was obtained from an original article or a replication protocol. The top row shows for each QMP type the proportions of each QMP ratio obtained, darker colours represent proportionally more QMPs, grey means modification did not occur for that measure. The bottom 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

A beta-regression model was used to test whether or not the difference in QMP ratio between original and replication was significant. Beta regression models are similar to other generalized linear regression models. Beta-regression is 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, 2010a). The beta regression model was implemented using the betareg function within the betareg package (Cribari-Neto & Zeileis, 2010b; Grün, Kosmidis, & Zeileis, 2012) with default parameters.

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 of indicated that this difference was significant (, 95% CI , , ). This result is in line with hypothesis 4.

### H5.

A logistic regression model was used to test the association between the QMP ratio in replication protocols and replication success. In line with hypothesis 5, results showed that a decrease in the ratio of QMPs in replication protocols significantly related to successful replication (, , 95% CI , , ).

### H6.

The relation between QMP ratios in the replication protocols and corresponding original articles were investigated using a beta regression model. Total QMP ratio in the original article was found to be significantly related to the total QMP ratio in the subsequent replication (, 95% CI , , ). This result provides evidence in favor of hypothesis 6. Figure 5 displays this relationship visually.

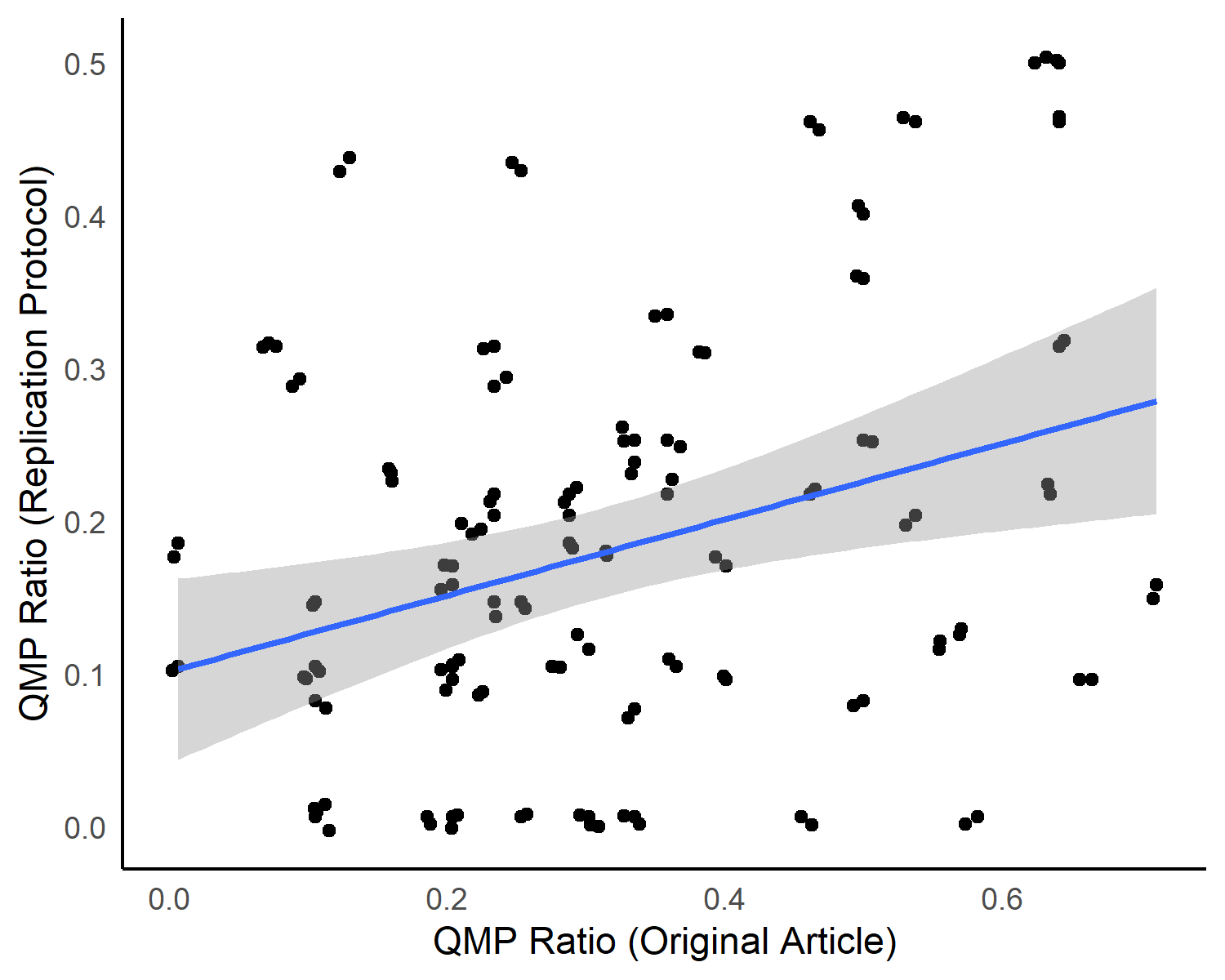


Figure 5: Scatterplot of original and replication total QMP ratio’s with linear regression line. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots in order to show the number of observations at points where multiple dots were present.

# Discussion

In this article, we analyzed the data, protocols, and related original articles from four Many Labs projects to assess reliability and measurement (reporting) practices. We additionally looked at how these features might affect credibility, in particular through replicability. Overall, we found that not all measures were sufficiently reliable, nor where measures always equally reliable in each setting in which they were used. Regardless, we did not observe that reliability was related to replicability as predicted. However, a relation with replicability was observed for QMPs in replication research. In turn, we found that QMPs in the replication were related to QMPs in the original research.

## Reliability

### Reliability varied within and between measures.

The measures in our sample varied in their reliability. 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, original research generally showed lower reliability than the average reliability in the replication samples, contrary to our hypothesis that there would be no difference.

One explanation for this discrepancy might be that measurement was more rigid in the Many Labs replications, possibly because the sharing of data would make them more open to scrutiny (Wicherts et al., 2011). Higher reliability would mean less noise around effect estimates, and in turn make those estimates more consistent. However, this relationship is complex and is further complicated by the fact that reliability is not a fixed aspect of a scale, but instead a feature of that scale within a sample (Cho & Kim, 2015).

Indeed, our results show that reliability of a measure is different 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 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. Regardless, our results do demonstrate an often overlooked aspect about of measurement 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 suprising 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, because lower reliability increases the noise in the data, then according to statistical theory the effect size estimates based on those data should decrease.

However, the relationship between reliability and replicability is complex. 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 would be expected. Furthermore, while lower reliability decreases effect sizes it can also reduce the observed true variance in the effects (Olsson-Collentine, Bakker, & Wicherts, 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 become even more complex when reliability itself is heterogeneous as was observed with several measures in our sample.

The complexity of the effect combined with our small sample of measures for which reliability could be calculated and for which the related true effect was likely not null may explain our findings. We likely did not detect that lower reliability would be associated with a decreased chance for successful replicability, because of a lack of statistical power.

The small number of relevant measures and reported reliabilities also indicate 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 similar investigations in the literature (Flake et al., 2022; Flake et al., 2017; Shaw et al., 2020). Reported reliabilities were so few, that the preregistered test for hypothesis 1 would not be informative. In particular, we found that replication research reported reliability coefficients less 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 been addressed in literature for a long time (Green, Chen, Helms, & Henze, 2011; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). Our study illustrates three potential reasons for why this problem has persisted. Firstly, the studies in our sample commonly made use of single item measures, an observation that was also made by Shaw et al. (2020). Calculating reliability for single-item measures is not as straightforward as for multiple item measures.

Secondly, our results indicate that there may be bias in reporting reliability coefficients. Reliabilities were reported in original research for those measures that obtained large 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. Researchers may be reluctant to report reliability for unreliable measures, thus lowering the prevalence of reported reliabilities. This bias may cause researchers to be overly optimistic with regard to 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, Duval, Tweedie, Abrams, & Jones, 2000).

Thirdly, when the replication protocols were written, no measurement based on these protocols had taken place. As a result, the data the replicating researchers could have calculated reliability and validity from was not yet available. This would explain why there was little reliability and validity evidence reported in replication protocols. However, if reliability and validity evidence are not being consulted before replicating research, it means that unvalidated measures are used.

Beyond reporting of reliability and validity evidence, we also investigated other measurement reporting practices recorded using QMPs. We found that 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). 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. One example of how this may have reduced QMPs is the inclusion of a section in several protocols for 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 hypothesis 5 and 6, we found 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).

## Limitations & Future Research

Our study had several limitations. First of all, we encountered numerous challenges in operationalizing QMPs. This research used a novel way of operationalizing QMPs as a variable to assess measurement practices and their impact on replicability. However, this also means that it remains uncertain whether or not our operationalization accurately captured the concept. We already determined it necessary to revise our coding protocol to be more lenient with regard to more context dependent QMPs compared to the initial protocol. This revision was not trivial, as it alters the interpretation of our results.

This relates to one of the most prominent challenges in assessing QMPs. It is difficult to determine when a practice is questionable, and what impact the context has on this decision. For example, in our sample the constructs that were measured were less commonly defined in replication protocols than in original articles. However, one may argue that it is not the responsibility of the replication to define the construct, as it is meant to be equivalent to the construct assessed in the original article. In that case, not defining the construct may not be a questionable practice at all.

Another challenge in constructing a QMP variable was the way to treat practices that were not applicable to a particular measurement. We chose to make use of 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. However, 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 was not applicable. As a result, the relation between QMP ratios and measurement reporting completeness may have been substantially obscured.

The relation is further obscured because it is not clear how great a ratio of QMPs represents 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 set out in the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (Eds.) (2014) standards for measurement.

The second limitation is that our tests on reliability were largely 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, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012; Nunnally, 1978), and are more limited with respect to the type of reliability and validity evidence that can be determined (Shaw et al., 2020). Because of these issues, it could even be argued that the use of single-item measures is a QMP in many scenarios.

As a third limitation, it is important to consider whether or not replication protocols and research articles can be fairly compared in terms of QMPs. A study’s description within a replication protocol is generally shorter than an article. The description within the protocol may therefore lack the space needed to report on the measurement in full detail. There are two main reasons why we believe protocols and articles remain comparable. Firstly, articles are often also restricted in the amount of space they have available to devote to measurement (Gardiner, 2019). Secondly, in the revised protocol some information, such as example items, could also be reported only in supplementary materials and not be counted as a QMP. Still, future research may wish to include supplementary materials as part of the primary data further to get a more complete picture on the state of measurement reporting in both articles and protocols.

The fourth limitation is that we made use of only one way to define replication success. Replication success can be estimated and framed in multiple ways, including methods that do not make use of significance testing. Multiple ways to define replication were used in the RPP (Open Science Collaboration, 2015). The relationship that reliability has with different indices of replication success may differ.

Finally it is worth noting that we did not make use of a causal design in our study. Therefore, we cannot attach definitive causal directions to the relations observed in this study. However, it would be useful for understanding the potential negative impact of reliability and QMPs on replicability. To achive this, future research could randomly assign original studies in a large scale replication project to either needing to report all the relevant measurement information as noted down in Table 1 in Flake and Fried (2020), while other studies are not screened. The same would go for requiring the use of 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 not. Tremendous coordination and funding would be required 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. For example, Rohrer (2018) has suggested researchers make use of Directed Acyclic Graphs to investigate causality in (non-)experimental settings by systematically taking into account confounding influences on the effect. Such an approach may also help in untangling the complex relationship between reliability and replicability.

## Recommendations

Taking all the findings together, the following assessments will be relevant to consider for future research. 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 suggest that for multiple item scales establishing the validity becomes a community standard for credible research. For reliability coefficients we go even further to state that these should by default always be reported for any multiple item scale. Researchers should only be exempt from this default, if they provide motivated reasoning for why they should be exempt.

For direct replications to represent an as good as possible test of the credibility of a finding, it is important that the procedure and measurement of the original study can be mimicked. Therefore, researchers should make fully report on their measurement details, as specified for example in Table 1 in Flake and Fried (2020), or section 3.6 in American Psychological Association (2020) on measures and measurement. Any details that could not fit in the article can then be shared with potential replicators via 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. Not only is it crucial for an informative replication that the measurement is reliable or valid. The original study should also report the measurement details necessary to reconstruct the original measurement. Otherwise it may be futile to attempt a direct replication. In that case 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. 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, validity, and reporting of measurement in the Many Labs replications and associated original studies, we found that reliability and validity evidence was reported infrequently. Furthermore, QMPs that obscured important information needed to evaluate and reconstruct the measurement were common, especially in original research. QMPs were in turn related to the replicability of a finding. Results with regard reliability of the measures were less clear due to low sample size. However, for a number of measures reliability showed signs of significant variation between studies. Combined these findings illustrate the need for more concrete standards in measurement reporting, and that before attempting a replication of a study, researchers should consider the validity and reporting completeness of the measurement.

# 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.

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*.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. <https://doi.org/10.1126/science.aaf0918>

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>

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. (2010b). 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. (2010a). Beta regression in *r*. *Journal of Statistical Software*, *34*(2). <https://doi.org/10.18637/jss.v034.i02>

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>

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., … 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., … 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.

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>

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

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of american public opinion. *The Teaching of Contemporary Affairs: 21st Yearbook of the National Council of Social Studies*, 11–34.

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., … 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., … 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., … 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>

Li, S., Jiang, H., Yang, H., Chen, W., Peng, J., Sun, M., … 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>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … 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., 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* (pp. 97–146; B. B. Wolman, Ed.). Boston, MA: 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>

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*. Boston, MA: Posit Software, PBC. Retrieved from <http://www.posit.co/>

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., et al.others. (2020). *High replicability of newly-discovered social-behavioral findings is achievable*.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <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>

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>

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.

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>

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*. Evanston, Illinois: Northwestern University. Retrieved from <https://CRAN.R-project.org/package=psych>

Zenodo open data repository (CERN). (n.d.). Retrieved from [https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/ DataPortal/Zenodo](https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/   DataPortal/Zenodo)

# (APPENDIX) Appendix

# 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. In order 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 (Bates, Mächler, Bolker, & Walker, 2015) t.

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 or not 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 gives a test of hypothesis 2, as outlined above.

## Unidimensionality Exploratory Analyses

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

In order to present additional validity evidence alongside the analyses of Cronbach’s alpha and omega coefficients, the unidimensionality of each measure was investigated. 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 due to the fact that 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 uni-dimensional. 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.

Nevertheless, we present here the result of the analyses made under the assumpti on that the one score, one dimension heuristic holds, for the readers that are comfortable with this heuristic.

## Model

A single factor confirmatory factor analysis was implemented using the *fa* function from the *psych* package (William Revelle, 2023) 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 2. 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 analysed.

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, Keith, Zhu, & Chen, 2013). However, it should be noted that at least with respect to Many Labs 2, all of the 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. Using a similar model to the main hypothesis test with all unique combination of the five different QMP types being tested across original and replication data. 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.

It is partly this large number of tests, as well as the resulting lack of clear interpretability in the results that caused this analysis to be omitted from the main article. However, for completeness and transparency it is 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 5 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.

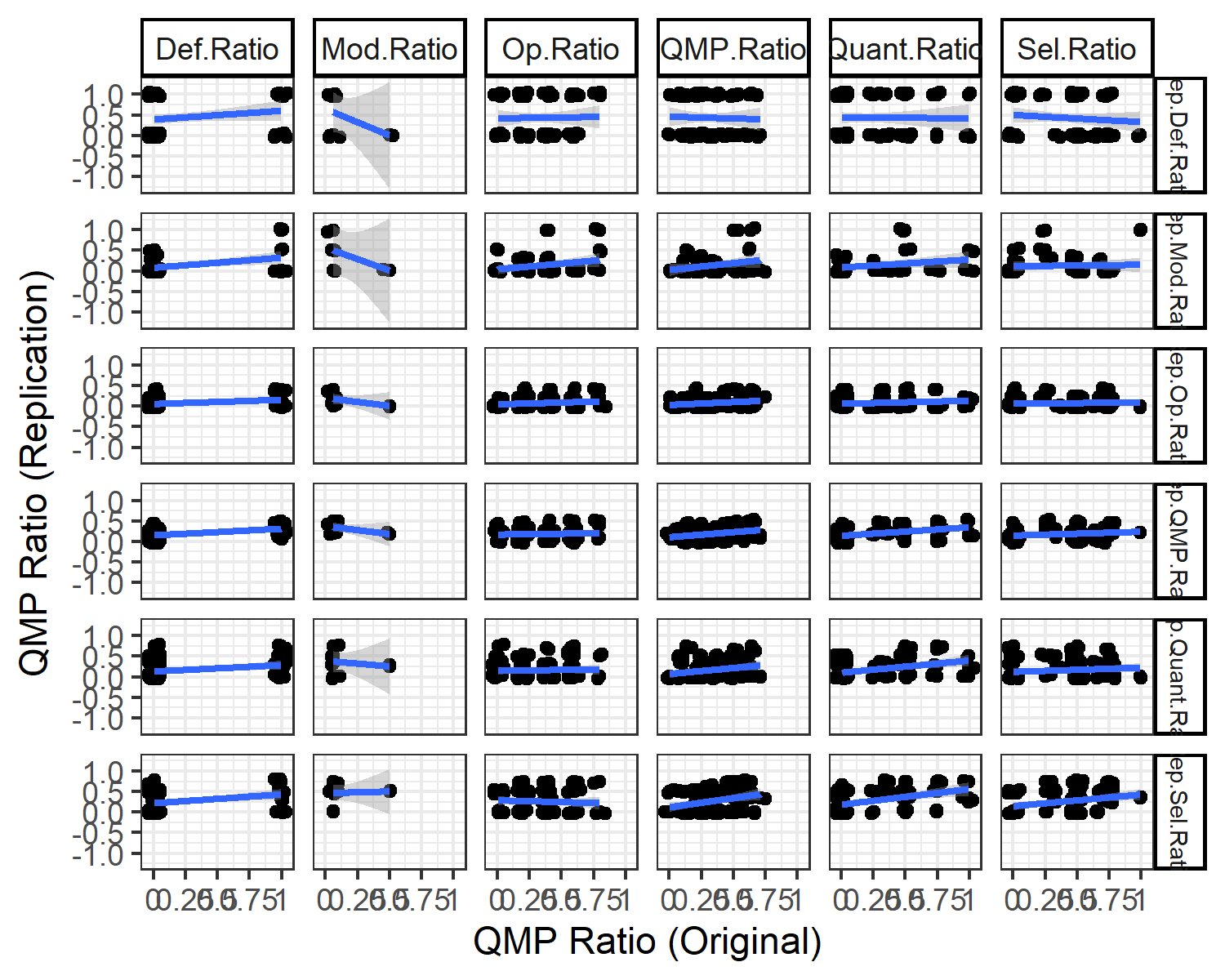


Figure 6: Scatterplot of original and replication QMP ratio’s per type with linear regression line. The abbreviations on the axes relate to the QMP categories described in Table 1: Def. is definition, Op. is operationalization, Sel. is selection/creation, Quant. is quantification, and Mod. is modification. 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 in order 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 of 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 5), 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.

# 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 published article for the sake of brevity, and because the sample size was not large enough to support multilevel modelling in most cases.

## Hypothesis 3, nested within replication set

### Explanation of Relevance.

In this data, replication attempts at different lab locations can be seen as nested within a replication set. In order 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.

In order 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. This result provides some additional evidence for the stability of hypotehsis 5.

For hypothesis 6, the sensitivity analyses using a random intercept logistic multilevel model resulted in non-significant results, which is contrary to the results found with the main test and is not in line with hypothesis 6 (revised protocol: , 95% CI , , ). This result places some doubt in the results found through the main test reported in the article.

# 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 why.

Omega was included since it has been argued to be a more informative measure of score reliability than alpha, while also providing validity evidence for the scale (Crutzen & Peters, 2017; Deng & Chan, 2017). Omega was calculated using the omega function in the psych R package (William Revelle, 2023). Default arguments were used in the function except the nfactors argument, which was set to 1.

### Models.

In order 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 (Bates et al., 2015).

The ICC was calculated similarly based on the omega as was done for Cronbach’s alpha in the article.

The analysis for H3 related the calculated average reliability coefficients for each replication set to whether or not the replication was successful when judged at meta-significance level of .

A 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 Appendix B.

### 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.

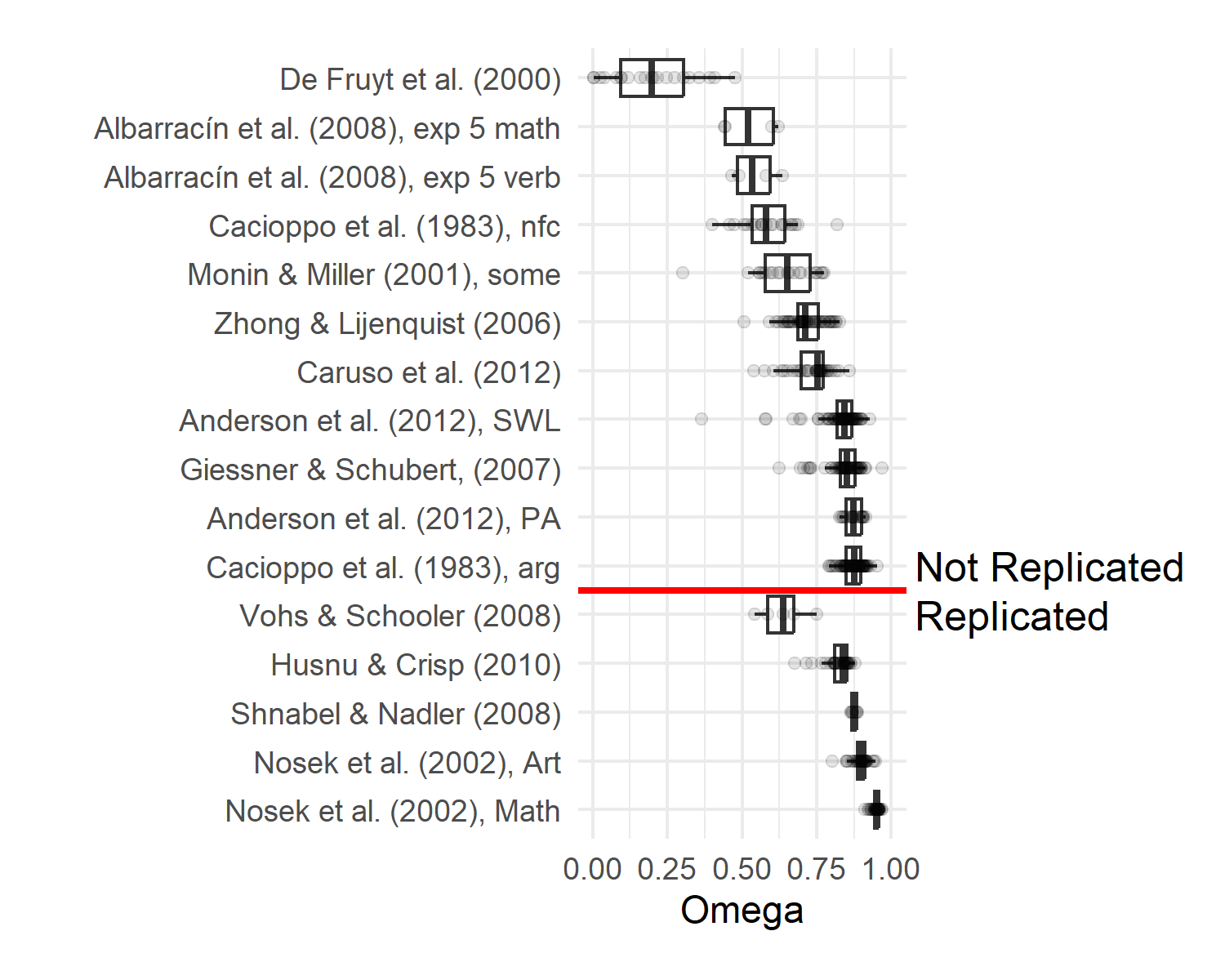
The calculated Omega coefficient did not significantly relate to replication success in the main logistic regression model. Similarly, non-significance was also observed in the multilevel random-intercept and random-slope models (random intercept: , 95% CI , , ; random slope: , 95% CI , , ).

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 for 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 2 from the article, with some studies being omitted as the omega coefficient function did not converge or run properly on those data. Furthermore, information with regard to 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.



# Initial Coding Protocol QMP Analyses

This appendix contains all of 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 to 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.

The relation between QMP ratios in the replication protocols and corresponding original articles were investigated using a beta regression model.

Appendix B 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 ratio’s across original articles and replication protocols based on the initial coding protocol.

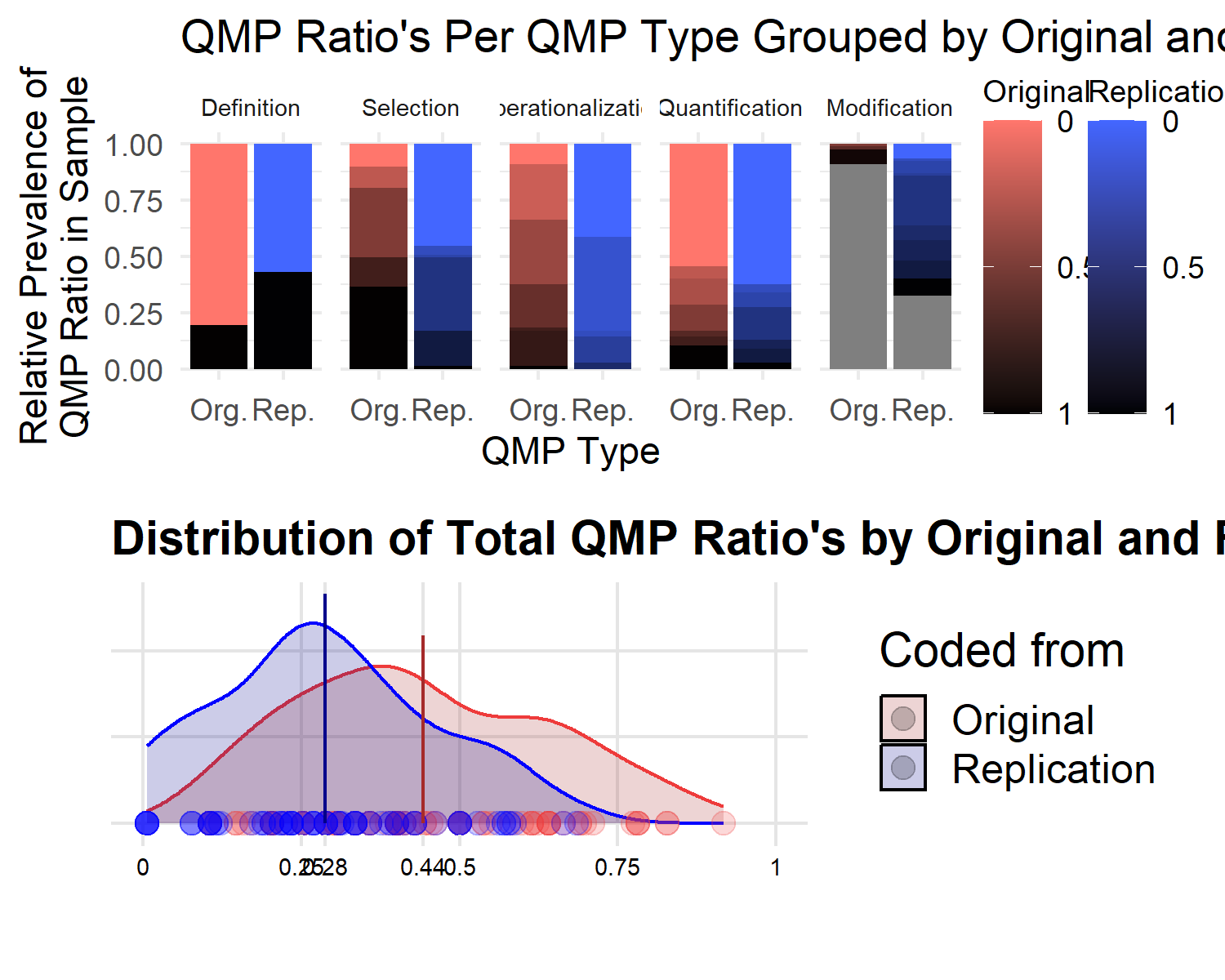


Figure 7: QMP ratio counts for each QMP Type and QMP total ratio distribution grouped by whether the QMP ratio was obtained from an original article or a replication protocol. The top row shows for each QMP type the proportions of each QMP ratio obtained, darker colours represent proportionally more QMPS, grey means modification did not occur for that measure. The bottom 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 panel of Figure 7 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% CI , , ). 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 8 displays this relationship visually.

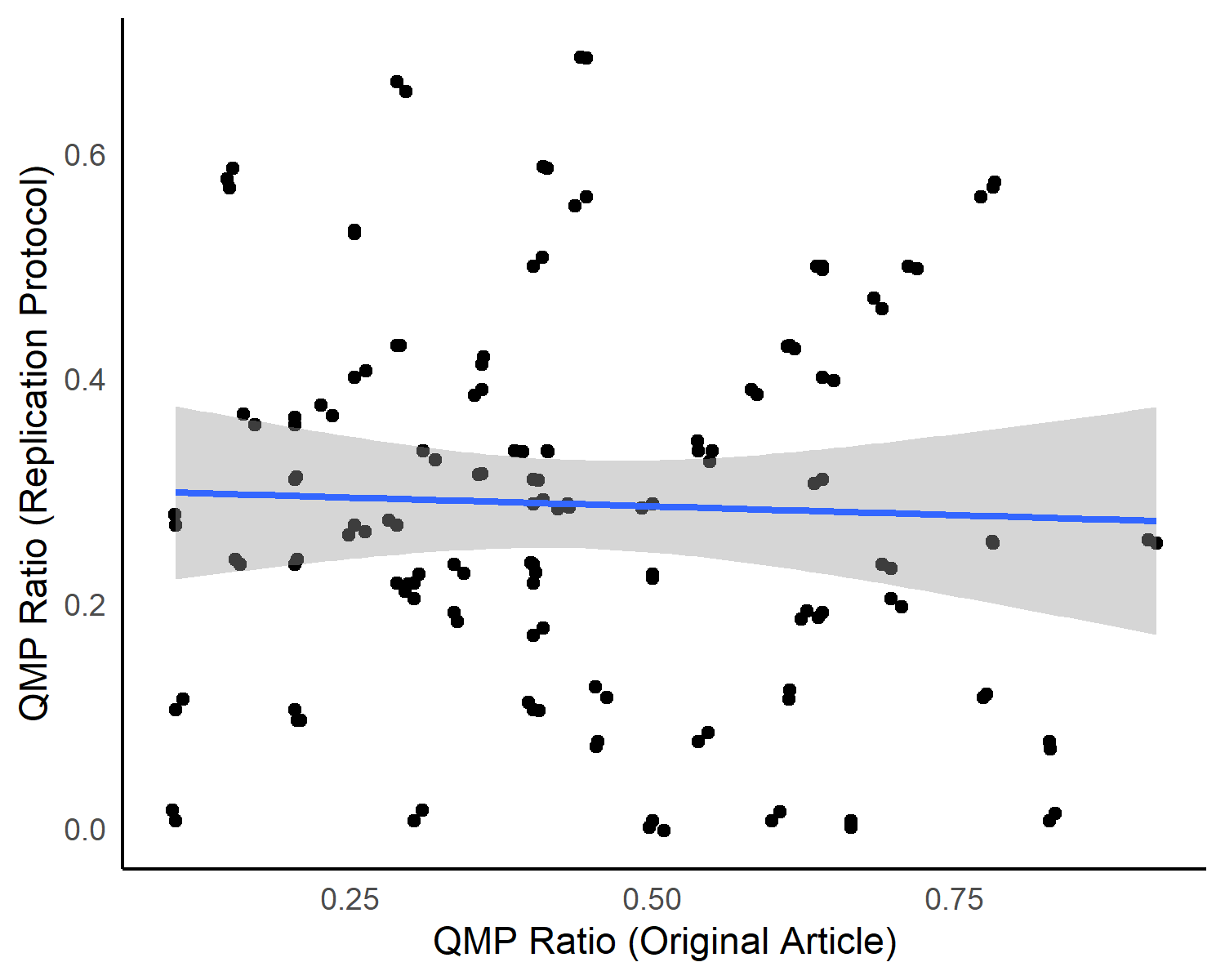


Figure 8: Scatterplot of original and replication total QMP ratio’s with linear regression line. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots in order to show the number of observations at points where multiple dots were present

Table 7 shows the relations between original and replication QMP ratios for each QMP type based on the initial protocol. Figure 9 shows the same relations displayed visually.

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.

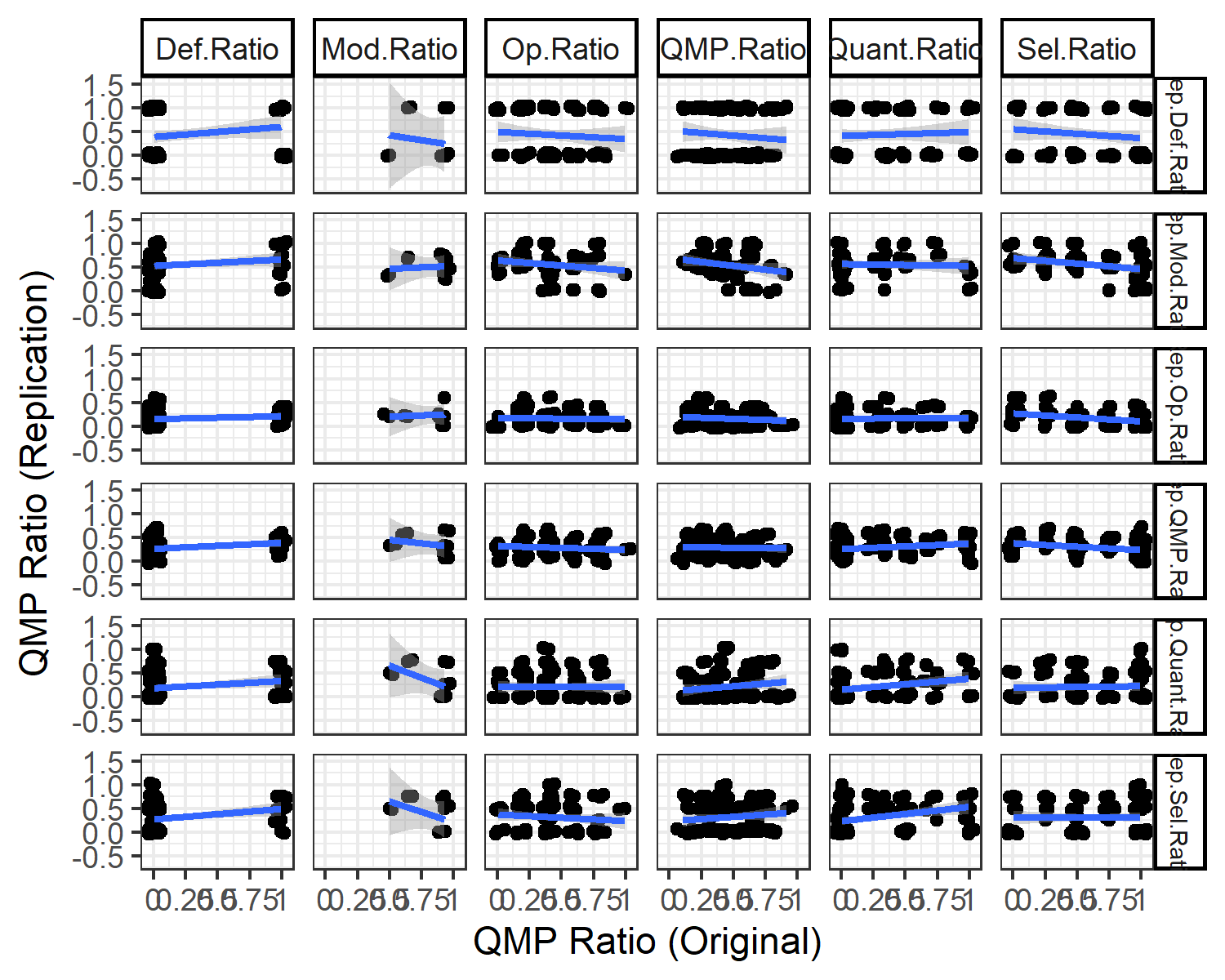


Figure 9: Scatterplot of original and replication total QMP ratio’s obtained using the intial coding protocol with linear regression line. Each dot in the figure describes the QMP ratio in that graph across both the original article and its replication protocol. Note: jitter was applied to these dots in order to show the number of observations at points where multiple dots were present

Now looking at the multilevel sensitivity analyses (see appendix B), only now using the QMP ratios based on the initial coding protocol.

Contrary to the results of the multilevel sensitivity model testing hypothesis 5 using QMP ratios obtained with the revised coding protocol, the relationship in the multilevel model was no longer significant for QMP ratios obtained with the initial coding protocol (, 95% CI , , ).

Similarly to the results of the multilevel sensitivity model testing hypothesis 5 using QMP ratios obtained with the revised coding protocol, the relationship in the multilevel model was not significant for QMP ratios obtained with the initial coding protocol (, 95% CI , , ).

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 was described as measuring 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 to make comparison between original and replication easier throughout the article. [↑](#footnote-ref-41)
2. The steps taken to access the replication protocol data from Many Labs 5 were not preregistered, and only became clear upon further investigation of the accessible information. The specific names and locations of the files that were used as the datasets can be found in the data\_search\_details.pdf supplementary materials of this article (<https://osf.io/gft46/%7D>). [↑](#footnote-ref-50)
3. An issue was encountered while trying to access Hyman and Sheatsley (1950). However, after contacting the first author of Klein et al. (2014) a suitable version of the article was obtained. [↑](#footnote-ref-53)