



Swiss Federal Institute of Technology Zurich

Seminar for
Statistics

Department of Mathematics

Master Thesis

August 2021

Stefan P. Thoma

Estimating Relevance within Replication Studies

Submission Date: August 31st 2021

Co-Adviser: Prof. em. Dr. Werner Stahel
Adviser: Prof. Dr. Martin Mächler

Abstract

This study investigates the application of Relevance as defined by [Stahel \(2021b\)](#) to three (Many Labs) replication studies. It starts with an introduction to the replication crisis in psychology and discusses how replacing p-values with Relevance could facilitate a more nuanced and reliable scientific process. The re-analysis using Standardized Mean Difference (SMD), log Odds Ratios (LOR), and the adjusted coefficient of determination R_{adj}^2 is then presented.

The relevant effect (SMD) of experiment three by [Albarracín et al. \(2008\)](#) failed to replicate with none of the eight attempts by [Ebersole et al. \(2020\)](#) producing relevant results. They further lacked precision to either establish equivalence between the groups nor to prove the original effect to be an anomaly. Re-analysis of the replication of [Schwarz et al. \(1985\)](#) by [Klein et al. \(2014\)](#) successfully produced relevant LOR at 15 out of 36 attempts. Although precision of the estimates varied, almost all point estimates were relatively close to the original effect. The four replication attempts of [LoBue and DeLoache \(2008\)](#) by [Lazarevic et al. \(2019\)](#) conclusively showed the effect of the experimental group on R_{adj}^2 to be negligible.

Relevance proved to be a widely applicable and intuitive measure which, unlike p-values, can embed evidence against *as well as for* the null-hypothesis and does not confuse precision with magnitude. Still further studies are needed to develop concrete procedures for establishing negligibility of effects.

Contents

1	Introduction	3
1.1	Replication crisis	4
1.2	Assessment of Success	4
1.3	Impediments of replicability	5
1.4	Relevance	9
2	Methods	11
2.1	Measures	11
2.1.1	Standardized mean difference (SMD)	11
2.1.2	(Log) Odds Ratio (LOR)	12
2.1.3	Adjusted R^2 (R^2_{adj})	12
2.2	Relevant directional replication	12
2.2.1	Power Calculation	13
2.3	Comparing effect magnitudes	14
2.4	Heterogeneity	16
3	Data	19
3.1	Replication projects	19
3.2	Studies	19
3.2.1	Albarracín et al. (2008) (alb5)	19
3.2.2	Schwarz et al. (1985) (schwarz)	20
3.2.3	LoBue and DeLoache (2008) (lobue3)	20
4	Results	23
4.1	Alb5: SMD	23
4.1.1	Effect magnitudes	23
4.1.2	Heterogeneity:	24
4.1.3	Comparison to ManyLabs	24
4.2	Schwarz: Log Odds Ratio	25
4.2.1	Differences in effect	25
4.2.2	Heterogeneity:	25
4.2.3	Comparison to ManyLabs	25
4.3	Lobue3: R^2_{adj}	27
4.3.1	Comparison to ManyLabs	27
5	Conclusion	29
	Bibliography	38
A	Tables	39
A.1	Alb5	39
A.2	Schwarz	40
A.3	Lobue3	42

List of Figures

1.1	Relevance categories. Figure reproduced from Stahel (2021b) with permission from W. Stahel.	8
2.1	Estimated Relevance of R-squared and adjusted R-squared for multiple predictors under H_0	13
2.2	RL.Power and Sig.Power over sample size.	15
2.3	Best case of ranges of confidence intervals for effect differences given the CI ranges of the original and replication effects.	16
3.1	Notation and structure of the data.	20
4.1	A: SMD estimates with CI and relevance threshold $\zeta = 0.1$. B: Difference between original effect size and each replication attempt.	24
4.2	A: Logit estimates with CI and relevance threshold $\zeta = 0.1$. B: Difference between original effect size and each replication attempt.	26
4.3	R2 estimates with CI and relevance threshold for the lobue3 study, $\zeta = 0.1$	27

List of Tables

A.1	Effect estimates and relevance of the alb5 replication analysis.	39
A.2	Alb5 effect magnitude comparison between replication attempts and the original study.	39
A.3	Effect estimates and relevance of the schwarz replication analysis.	40
A.4	Schwarz effect magnitude comparison between replication attempts and the original study.	41
A.5	R squared estimates and relevance of the lobue3 replication analysis.	42

Chapter 1

Introduction

This work builds upon a seminar paper and code by [Herger and Rogai \(2018\)](#).

Language of reproducibility: I will use the terminology suggested by [Stahel \(2021a\)](#) to differentiate concepts related to the vague idea of reproducibility. *Re-Assessment* describes the process to go through the exact same analysis with the same data of the original publication. Difficulties can arise through lack of transparency of analysis and the unavailability of the data. *Re-Analysis* uses the same data as the original publication but develops a new analysis pipeline. This can be required if there were serious problems in the original analysis, e.g. lack of multiple testing adjustment, or if different methods may give answers to new questions. *Replication* requires the realization of a new experiment generally on a new sample which aims to reproduce the effect of the original study. A replication can either confirm the original study, or fail to do so.

Replications can be either *exact / direct*, can go a step further and aim to investigate the *robustness* of an effect by varying the experiment, or *generalize* an effect by changing the experiment by modifying either study design and/or the data generation process ([Stahel, 2021b](#)). A *conceptual replication* has as its aim the same idea of effect, but may develop a completely different experiment.

This study represents a re-analysis of series of direct replication studies.

Scientific method: Traditionally, the frequentist scientific method involves setting up a null-hypothesis (H_0) and an alternative hypothesis (H_1) then gathers data to either reject H_0 and accept H_1 or fail to reject H_0 and hence accept H_0 . When making this decision two errors are possible: Either rejecting H_0 although it is true (type I error, its probability is α), or not rejecting H_0 although it is false (type II error, its probability is β). Conventionally, probability of a type I error is fixed at 5% and sample size is chosen to ensure 80% power, where power is defined as $\text{Power} = 1 - \beta = 1 - P(\text{type II error})$ and in practice requires specifying the presumed effect size under H_1 .

Hierarchy of evidence: The quality of scientific evidence of a study can be judged by its internal and external validity. Internal validity relates to the study design, informally measuring how well the studies conclusions are supported by its theoretical arguments and empirical evidence. Randomized control trials are often thought of as the gold standard of experimental design when it comes to internal validity and are a very popular design choice for experimental psychologists. External validity is a broader quality which builds on internal validity and describes the degree to which the findings of a study can be

generalized to other populations, situations, and points in time. The external validity of results can be assessed by conducting replication studies designed to confirm or reject the findings of an original study. Even higher on the hierarchy of scientific evidence would be series of replication studies for a specific effect, or meta-analyses often covering more general concepts.

1.1 Replication crisis

Already in 2005 (2005) raised concerns that most published research findings are false. Around 2010, psychological researchers noticed that attempts to replicate established effects failed much more often than a type I error of 5% might suggest. As explained by Ioannidis (2005) and Gibson (2021) the probability of a false discovery depends not only on α but also on the prevalence (or prior probability) of an effect. One of the first large scale systematic investigation (led by Brian Nosek) found that out of 100 attempts to replicate an experiment published in one of three respectable journals only 36% to 37% of replications succeeded (Collaboration, 2015). This phenomenon had been dubbed the *replication crisis* and redirected more attention and resources of the psychological community on such meta-scientific topics (Fletcher, 2021).

Spanning from 2014 to now, there were a handful influential large scale replication studies where multiple studies were replicated by different research groups. The first of the so called Many Labs studies (see Stroebe, 2019, for a more detailed review) aimed to investigate variation of replicability in psychological studies under different conditions Klein et al. (2014). They found that ten of the thirteen chosen studies replicated relatively independent of the condition (e.g. lab vs. online, Stroebe, 2019). The Many Labs two project aimed to estimate the heterogeneity of effect sizes. While around 14 out of 28 studies successfully replicated, heterogeneity depended more on particular effects and less on conditions and varied between studies Klein et al. (2018). The third Many Labs project only replicated three out of ten studies (Stroebe, 2019). Klein et al. (2019) tested in the Many Labs 4 study whether involving the original authors in the experiment protocol might impact replication success. A very useful endeavor to verify that a replication is conducted in the same spirit as the original study. It could well be, for example, that the idea of an original paper could not be replicated using the exact same material as was used some years ago, while a conceptual replication might still work well (for a hypothetical example see Stroebe, 2019).

Extending the replication crisis, Yarkoni (2019) paints an even worse picture of the state of psychological research by calling out its low external validity and generalizability. He criticizes both the transition of vague concepts to very specific and mathematical models and the transition back from the model inference to general psychological insights (Yarkoni, 2019). Only conducting experiments at a much larger scale with as many variations as possible – included as random effects – could warrant the arguably already claimed generalizability (Yarkoni, 2019; Lakens, 2020)

1.2 Assessment of Success

To judge whether a replication was successful or not we must have a proper definition of a successful replication. As of yet, there is no universally agreed upon definition of replication success. In general, replications aim to replicate studies that did find an effect,

not replicate zero-effect studies. According to Bonnett (2021), there are two ways to judge a replication as successful: First, if the replication effect is significant and points in the same direction as the original effect (*directional replication*). Second, if we can establish an equivalence between the original effect and the replication effect (*strong effect size replication*).

Comparing effect size magnitudes really only makes sense when both experiments conducted the experiment and measured the variables in the same fashion. To establish equivalence Bonnett (2021) suggests specifying a *range of practical equivalence* (ROPE, from $-\zeta$ to ζ). We would accept two effect sizes (θ_1 and θ_0) as practically equivalent if the difference of effects lies within the ROPE ($-\zeta < \theta_0 - \theta_1 < \zeta$). There are alternative methods to establish equivalence, e.g. the *two one sided tests* (TOST) procedure or the bayes factor (Anderson and Maxwell, 2016, Linde et al. (2020)).

While a successful replication confirms the effect found in the original study either qualitatively or quantitatively, interpreting a failed replication is not as straightforward. The framework developed by Bonnett (2021) establishes four types of evidence against the original study effect (nonreplication evidence). The first is the *directional nonreplication*, where both effects are significant, but point in the opposite direction. Number two and three offer evidence of differences in effect size. The *strong effect size nonreplication* is when the confidence interval of the effect difference is completely outside the ROPE. For *weak effect size nonreplication*, we would reject the null-hypothesis of zero difference in effect ($H_0 : \theta_1 - \theta_0 = 0$). *Null effect size nonreplication* is the case when the replication effect is not significantly different from zero and its CI is within the bounds of $-\zeta$ and ζ . Any other constellation of results would accordingly be categorized as inconclusive. Maxwell et al. (2015), Anderson and Maxwell (2016), and Bonnett (2021) emphasize that failure to produce a significant effect in the correct direction does not necessarily negate the existence of the effect, nor does it indicate that the replication study was faulty.

Another method of assessing success in replication studies are *meta-analytic* procedures where one would update effect estimates after each replication to include all data subsequently collected (Fletcher, 2021). Such procedures can lead to more generalizable conclusions but have to account for heterogeneity of effects and publication bias (Anderson and Maxwell, 2016; Bonnett, 2021; Fletcher, 2021).

1.3 Impediments of replicability

Study design: A replication study can be insufficiently powered and therefore fail to find an effect (Maxwell et al., 2015). On the other hand, overpowered studies may also be problematic when judging success by the directional replication criterion because increased precision may produce spurious but significant results (Stahel, 2021b). Failure to replicate can also stem from badly designed replication studies that fail to capture the concept as intended by the original authors. Study design problems generally have to be dealt with before data collection. Three ways to avoid problematic study designs of a replication study are: Conducting a power analysis in the planning stage (Brandt et al., 2014), cooperating with the original authors (as done by Ebersole et al., 2020), and publishing a registered report where the experimental protocol (and the analysis plan) are peer-reviewed before data collection starts (Nosek and Lakens, 2014). Support for registered reports is steadily rising (see Center for Open Science, 2021, for more resources and a list of 294 participating journals).

Heterogeneity of effect: Even with all controllable variables held constant, effects can vary between replication attempts. Reasons for heterogeneity include differences in samples, a temporally varying effect, other hidden confounders, or stochastic processes (Kenny and Judd, 2019). When looking at a number of replication studies or conducting a meta-analysis, mixed effects models (MEMo) can (and should) be used to estimate effect heterogeneity. Estimated effects can depend heavily on study-specific circumstances and the precise stimuli and measurements used. Any variation in study procedure could introduce more heterogeneity (Yarkoni, 2019).

If heterogeneity is ignored a single large sample study can produce a very precise effect estimate. Unfortunately, if effect heterogeneity is present such an effect can be tied to the specific sample (and circumstances) used in the study. Thus, the confidence interval around the estimate would be too small to contain the true population level effect at the specified precision (Kenny and Judd, 2019). Ahn et al. (2012) proposed a method to appropriately increase the confidence interval in meta-analyses if the magnitude of heterogeneity is known. Heterogeneity is not known a priori but would have to be estimated by e.g. conducting at least five replication attempts (Stahel, 2021a; Hedges and Schauer, 2019).

Questionable research practices can stem from both malicious intent and lack of understanding of statistical principles. While fraudulent science by social psychologist Diederik Stapel arguably planted the first seed of distrust of psychological research in 2011 replications are not designed to uncover frauds (although they occasionally have, see Stroebe, 2019). A recent study by Francis and Thunell (2020) developed a promising framework to assess whether the findings of a study are *too good to be true* by estimating the probability of the acquired results given the discovered effect sizes and the power of the study.

Some common questionable research practices include: *P-hacking*, which is trying out different statistical tools until the desired effect becomes significant (Head et al., 2015). When analyzing data, researchers have to decide between various methods, e.g. outlier removal strategies. This choice is referred to as *researchers degree of freedom*. *HARKing* (hypothesizing after the results are in) is a data-driven analysis strategy which develops post-hoc theories that justify the results and present the findings as if they were theory driven. Essentially, if data on enough variables is collected finding some significant effect becomes trivial (Kerr, 1998). *HARKing* combines a lack of multiple testing adjustments with questionable reporting habits as it passes off exploratory findings as confirmatory results.

Pre-registration of analysis counters both p-hacking and *HARKing*. A pre-registered analysis plan ensures that a researcher cannot choose the method of analysis based on the results produced. A more detailed analysis plan leads to less researchers degree of freedom. Such a plan requires hypotheses to be specified before data are in. While pre-registration may seem limiting, it does allow for unregistered exploratory analyses as long as findings are declared as such.

At the moment there is no one agreed upon standard for pre-registrations. The Center for Open Science (OSF, 2016) currently offers ten registration templates, three of which specifically for replication projects. Pre-registering on OSF is very easy and accessible. Unfortunately, as of yet, there is no systematic database that includes all replication attempts and their outcomes. There is also no system following up on pre-registered studies making sure that results are published. This theoretically leaves the door open

for fraudulent activities, e.g. where many hypotheses could be pre-registered and only the successful one reported.

Pre-registration is much more strict in medical sciences. For example, the largest registration website *ClinicalTrials.gov* involves more bureaucracy and requires results to be submitted within a specific time frame (Wikipedia, 2021a). In psychology, such regulations may not be necessary as the stakes are somewhat lower but it illustrates that there is room for improvement.

Publication bias describes the fact that significant results are more likely to be published than null-results Martin and Clarke (2017). This is due to both the journals favoring new and positive findings, and the researchers lack of effort to publish null-results (a phenomenon known as file drawer effect, Rosenthal, 1979). Rosenthal (1979) already argued in 1979 for the acceptance and importance of published null-results. Because when various researchers investigate a non-existent effect, assuming the statistical testing procedure was applied properly, we would expect around 5% of experiments to show significant results. If those results get published but the null-results do not, a review of the the scientific literature would provide misleading evidence in favor of the existence of the effect.

Researchers and publishers share responsibility for publication bias. To counter it, the scientific community must grow more acceptable of null-result publishing. As registered reports focus on the relevance of a scientific question (and the quality of the experimental design), *not* on study outcomes, they are a viable measure against publication bias. Understandably, a journal filled with null-effects may not be the most interesting read so alternative modes of result-publication, like well organized public databases, may be of use.

Problems with the p-value are manifold and start already with its convoluted and un-intuitive interpretation: It represent the probability to receive the given data or something more extreme given that the null-hypothesis is true. Unfortunately, it is often misinterpreted as either the probability of the null-hypothesis being true given the data, or as the probability that the outcome happened purely by chance (Colquhoun, 2017).

P-values are a measure of precision and do not directly translate to effect sizes. Abandoning p-values in favor of CIs would give a more nuanced understanding of an effect both about precision and effect magnitude. Cumming (2008) argues that especially for replications researchers should rely on CIs instead of NHST. However, using CIs for dichotomous decisions (e.g. accepting H_1 when CI does not include 0) is equivalent to NHST significance testing. Nowadays, often both are supplied.

A number of researchers proposed to simply reduce α to 0.005 to increase the evidential burden on researchers (although only for new discoveries, not replication studies Benjamin et al., 2018). Independently, Held (2019) also argued (in lieu of a full bayesian approach) for a decrease of α based on bayesian prior estimation. A very tantalizing solution that would supposedly “immediately improve the reproducibility of scientific research in many fields” (Benjamin et al., 2018, pp. 1) without introducing any new concept. However, calibrating α has been widely criticized for moving the goalpost without fixing underlying issues (and can actually make things worse Williams, 2019).

P-values are often associated with the importance of an effect although they are merely a measure of precision,. Psychological research historically treated p-values as the decisive

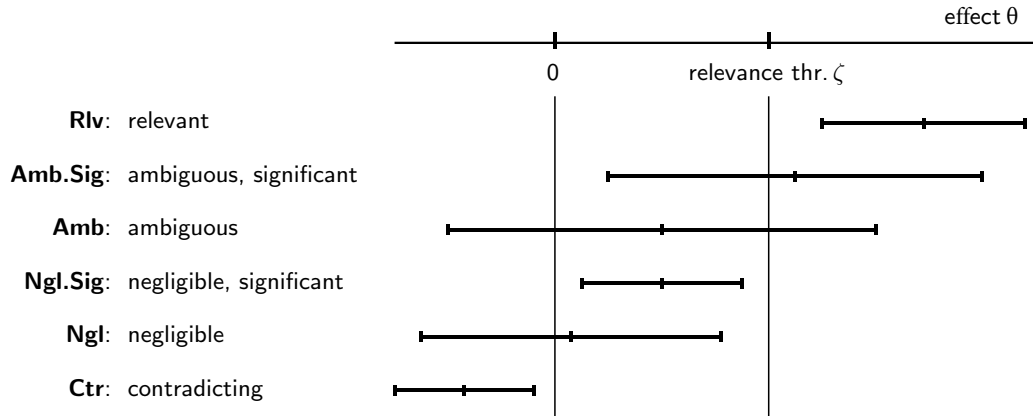


Figure 1.1: Relevance categories. Figure reproduced from [Stahel \(2021b\)](#) with permission from W. Stahel.

factor to evaluate whether an effect had been uncovered by an experiment ($p < 0.05$, success) or not ($p > 0.05$, failure). There are two related problems with such a procedure. First, there is intrinsically a lack of nuance with such a polar decision where $p = 0.048$ is treated very similarly to $p = 0.00002$, although the latter would generally be the preferred outcome of a study. Further, as the p-value ranges between 0 and 1 it quickly loses resolution as certainty increases and $p \rightarrow 0$. Second, it completely ignores the question of practical relevance, as the p-value does not directly relate to the magnitude of the effect. If there is *any* effect, uncertainty decreases - and the p-value with it - whenever our sample increases. This is problematic as an overpowered replication study would likely lead to a significant result whether or not there really is any effect.

What do we mean with an effect? In practice, detecting whether an effect differs from zero is often not a meaningful question. This is captured by the *zero hypothesis testing paradox* ([Meehl, 1967](#); [Stahel, 2021b](#)): It will almost never be the case that two groups are exactly the same with respect to some measurable variable. When we take a large enough sample of two groups we will almost certainly get a significant difference between the two samples even when our intervention or experimental condition did not have any impact on the dependent variable. Instead of asking whether our experimental condition has a (relevant) impact on the dependent variable, the p-value reduces the question to whether our precision was high enough to make the difference nominally significant ([Stahel, 2021b](#)).

The second generation p-value relates to the overlap of an effect interval to a null hypothesis interval ([Blume et al., 2019](#)). Importantly, it introduces thresholds of scientific relevance. The second generation p-value was designed to look and work like a regular p-value. While this may ease wide adoption in the scientific community it comes with some drawbacks ([Stahel, 2021b](#)): The coarseness of inference of the p-value remains; Either we reject or we accept, all nuance is lost. It also remains scaled between 0 and 1, thus limiting its power to convey very strong evidence of an effect.

1.4 Relevance

Relevance (Rl) defined by [Stahel \(2021b\)](#) is a model parameter that relates a (standardized) effect size to a pre-specified relevance threshold (ζ). The range $[-\zeta, \zeta]$ corresponds to the null hypothesis interval of the second generation p value. In the best case, ζ is chosen by the researcher and represents the minimal effect size considered to be scientifically relevant. Analogous to estimating an effect θ and its confidence interval CI_θ we have estimated Relevance (Rle) and its corresponding CI defined as:

$$Rle = \frac{\hat{\theta}}{\zeta}; \quad CI_{Rl} = \frac{CI_{\hat{\theta}}}{\zeta}$$

For simplicity a directed hypothesis is discussed. We speak of a relevant effect when Rl is larger than one, where the scale of Rl is in units of ζ . The null hypothesis can then be expressed with respect to the relevance parameter, $H_0 : Rl < 1$; $H_1 : Rl > 1$. The constructed CI around Rle ranges from secured relevance (Rls) to potential relevance (Rlp). We would classify an effect as relevant when $Rls > 1$.

By requiring a minimal relevant effect size the two strongest critiques of the p-value are answered: Obtaining a relevant effect by increasing the precision when the two populations of interest differ only slightly becomes very unlikely.

It further poses the more scientifically interesting question *is there a relevant effect*, rather than *is my sample size large enough to detect any difference between two groups*. The quote by [Tukey \(1962\)](#) summarizes the difference best:

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

Defining a relevance threshold larger than zero, inference is naturally more conservative compared to null hypothesis significance testing (*NHST*). However, because Relevance depends on the magnitude of the effect and not merely on the precision of the estimate, unlike decreasing the nominal p-value, Rle is largely unaffected by increased sample size. Compared to the second generation p-value, Rl offers a much more nuanced appraisal of evidence with (at least) six inferential categories of an estimate, see figure 1.1.

Specifics of replication Similar to [Bonett \(2021\)](#) there are generally two criteria of a successful replication study within the Relevance framework. One: Does the replication study lead to the same conclusion as the original study [Stahel \(2021a\)](#)? In general, replication studies aim to replicate an originally observed effect. If the original effect was relevant, a successful replication should produce a relevant effect as well. For ambivalent but significant (Amb.Sig) original effects expectations of a successful replications are not immediately clear. If a replication attempt has high power and the original study was estimated to be relevant (amb.sig: $Rle > 1$, $Rls < 1$) we would expect the replication to yield amb.sig or, even better, a relevant result. However, if for the original study $Rle < 1$, then even a highly powered study should not produce a relevant outcome.

Two: Are the effect sizes compatible or similar in a quantitative sense [Stahel \(2021a\)](#)? Another way to approach the question is to ask whether the observed effect in the replication study is equivalent to the effect observed in the original study. In the relevance framework one would construct a confidence interval around the difference of the effects and apply the relevance categories as defined in [Stahel \(2021b\)](#). Instead of aiming for a

relevant effect, in this case a successful replication would show a negligible difference in effects where both confidence limits are within the relevance thresholds around 0. This is very similar to the concept of equivalence testing [Lakens \(2017\)](#).

The present study aimed to re-analyze three multi-lab replication studies using different effect measures under the context of relevance. The three studies were borrowed from the two Many Labs projects by [Klein et al. \(2014\)](#) and [Ebersole et al. \(2020\)](#). In a first step, I will apply the relevance procedure to each replication attempt. Then, I will compare the magnitude of the effects of the replication attempts with the original study. I will further have a brief look at the meta-analytic solution, again applying relevance. The results are then compared to the results of the Many Labs project focusing on how Relevance can improve upon classic NHST inference.

Chapter 2

Methods

Methods described in this section were implemented in the package `ReplicationRelevance` specifically for this analysis (Thoma, 2021).

2.1 Measures

2.1.1 Standardized mean difference (SMD)

As the name suggests, the SMD is a measure for the difference in means between two groups. The difference is then standardized by the pooled standard deviation of the outcome variable (Y) per group (G).

$$SMD : \hat{\theta} = \frac{\bar{y}_{G1} - \bar{y}_{G2}}{\widehat{sd}_{\text{pooled}}}$$

Where the pooled standard deviation for two groups is:

$$\widehat{sd}_{\text{pooled}} = \sqrt{\frac{(n_{G1} - 1) \widehat{sd}_{G1}^2 + (n_{G2} - 1) \widehat{sd}_{G2}^2}{n_1 + n_2 - 2}}$$

Stahel (2021b) suggests a threshold of $\zeta = 0.2$ representing a small effect size. If there is a small effect size present, I would like to be able to classify it as relevante. Therefore, I chose to use $\zeta = 0.1$ as Relevance threshold.

We can calculate a (Wald) CI as:

$$\hat{\theta} - z_{\gamma} \cdot \text{se}(\hat{\theta}) \text{ to } \hat{\theta} + z_{\gamma} \cdot \text{se}(\hat{\theta})$$

On standardization: It often makes sense to provide the non-standardized effect as well. However, as the default Relevance threshold is defined for standardized parameters, Relevance generally relates to the standardized effect. If you want to work with non-standardized effects, e.g. if the scale of the dependent variable is very intuitive, you can still take the default relevance threshold and multiply it with $\widehat{sd}_{\text{pooled}}$. For multiple replication attempts this will result in varying relevance thresholds on the original effect scale as each will depend on the attempt specific $\widehat{sd}_{\text{pooled}}$.

2.1.2 (Log) Odds Ratio (LOR)

The Odds Ratio is a widely used metric for logistic regression models [Bland and Altman \(2000\)](#). As both predictor and dependent variable were categorical the OR is free of scale and thus not in need of standardization. Although the OR more easily interpretable, the LOR is approximately normal distributed and more convenient to work with. The LOR corresponds to the beta coefficient (β_1) of the logistic regression model:

$$P(Y = 1 | X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

Because LORs are approximately normal distributed *wald* CIs are used.

2.1.3 Adjusted R^2 (R^2_{adj})

The coefficient of determination (R^2) is generally defined as $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ ([Wikipedia, 2021b](#)). One can judge the predictive value of parameters based on the amount R^2 increases by adding parameters to a model. As visible in [2.1](#), R^2 will increase with additional predictors even when they have no impact on the dependent variable.

Alternatively, one can adjust the R^2 to account for the number of predictors used in the model. That way, the estimated R^2_{adj} will vary around 0 even for a large number of independent predictors (see [2.1](#)). The R^2_{adj} from the `stats` package is used for the analysis of the linear models ([R Core Team, 2021](#)):

$$R^2_{adj} = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

The CIs in figure [2.1](#) are simulation based. For the MEMo the $R^2_{F_{adj}}$ for the variance explained by the fixed effects is used, as proposed by [Zhang \(2021a\)](#) and implemented in the `rsq` package ([Zhang, 2021b](#)).

Of interest is the increase of R^2_{adj} for an arbitrary number of additional predictors. A relevant increase in R^2_{adj} would be:

$$Rl : \frac{R^2_{adj_{m1}} - R^2_{adj_{m0}}}{\zeta} > 1$$

Where $m1$ is a more complex model and $m0$ is a less complex model with fewer predictors. The CIs around the differences in R^2_{adj} were bootstrapped using the normal approximation. This was computationally expensive but allows easy implementation of many other measures into the `ReplicationRelevance` package source code in the future. As a decrease of R^2_{adj} is not meaningful, the strongest evidence against H_1 would be a negligible R^2_{adj} .

[Stahel \(2021b\)](#) lists the *drop effect* as a suitable measure for such a question. Reduction of R^2_{adj} was chosen to stay closely comparable to the results of [Lazarevic et al. \(2019\)](#).

2.2 Relevant directional replication

The effect size and CI was calculated for each replication attempt. A replication attempt was successful in directional replication if the total confidence interval of the effect estimate was larger than the Relevance threshold.

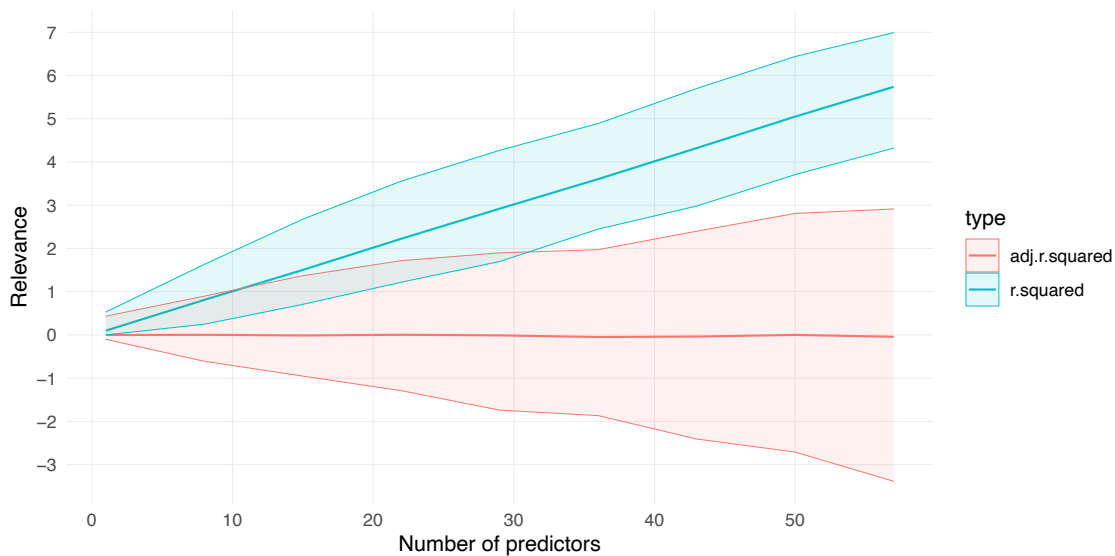


Figure 2.1: Estimated Relevance of R-squared and adjusted R-squared for multiple predictors under H_0 .

2.2.1 Power Calculation

The power of a study is traditionally defined as the probability of discovering a significant effect given a certain effect size and a prespecified sample size (*Sig.Power*). Applying this logic to the concept of Relevance, we define Relevance power (*Rl.Power*) as the probability of finding a *relevant* effect given a certain effect and sample size.

In the one sample z-test (where *sd* is known) the analytic solution to *Sig.Power* is straightforward (following the vignette *one-sample-z-test* by [Hayes and Moller-Trane \(2019\)](#)). Because we typically have a directed hypothesis in replication studies the following example is directed.

Sig.Power

$$H_0 : \mu = \mu_0; \quad H_1 : \mu = \mu_1 > \mu_0$$

The Z-statistic is as follows:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

To calculate power, we need to specify an effect size for the alternative hypothesis, $H_1 : \mu = \mu_1$. We can then calculate the probability to reject H_0 , given H_1 is true:

$$P(\text{reject } H_0 \mid \mu = \mu_1) = P\left(\bar{x} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right)$$

Now under H_1 , $\bar{x} \sim \mathcal{N}(\mu_1, \frac{\sigma^2}{n})$, thus:

$$\begin{aligned} P\left(\bar{x} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) &= P\left(\frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \\ &= 1 - P\left(Z < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \end{aligned}$$

RI.Power

With a relevance threshold we re-specify $H_0 : \mu < \mu_0 + \zeta$ $H_1 : \mu = \mu_1 > \mu_0 + \zeta$ or alternatively $H_0 = Rl < 1$, $H_1 = Rl > 1$ and get the probability to reject H_0 given H_1 as:

$$\begin{aligned} P(\text{reject } H_0 \mid \mu = \mu_1) &= P\left(\bar{x} > \mu_0 + \zeta + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) \\ P\left(\bar{x} > \mu_0 + \zeta + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) &= P\left(\frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 + \zeta + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > \frac{\mu_0 + \zeta - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \\ &= 1 - P\left(Z < \frac{\mu_0 + \zeta - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \end{aligned}$$

Using these formulas, figure 2.2 compares both RI.Power and Sig.Power for different combinations of μ_1 and sample sizes, given $\zeta = 0.1$. Two observations are apparent: As expected, RI.Power is generally more conservative than Sig.Power. This difference is especially obvious for the cases where $Rl \leq 1$ and $\mu_1 > 0$ where Sig.Power approaches 1 as $n \rightarrow \infty$ but RI.Power stays close to 0. At $Rl = 1$, we reject H_0 in only 5% of the cases. As soon as $Rl > 1$, RI.Power shows the same exact pattern at μ as Sig.Power showed at $\mu - \zeta$.

Estimating RI.Power: For more complex examples it is easier and more versatile to compute RI.Power and Sig.Power by simulation. This is done here for the SMD and the OR example both in the fixed and in the MEMo by supplying the original predictor values and subsequently simulate 1000 times based on the (standardized) effect size of the original study. A power estimate is not provided for the adj. R^2 example and for the effect size differences.

2.3 Comparing effect magnitudes

A comparison of effect magnitudes answers the question whether two effect sizes are equivalent or whether one is relevantly larger than the other. To estimate the difference (Δ) between two effects one should first work with unstandardized effect sizes and only standardize after based on the pooled sd . When standardized coefficients are compared, the effect size depends on the estimated sd of each replication attempt. So standardized coefficients could vary even if the unstandardized coefficients were exactly the same. That is not an attractive property.

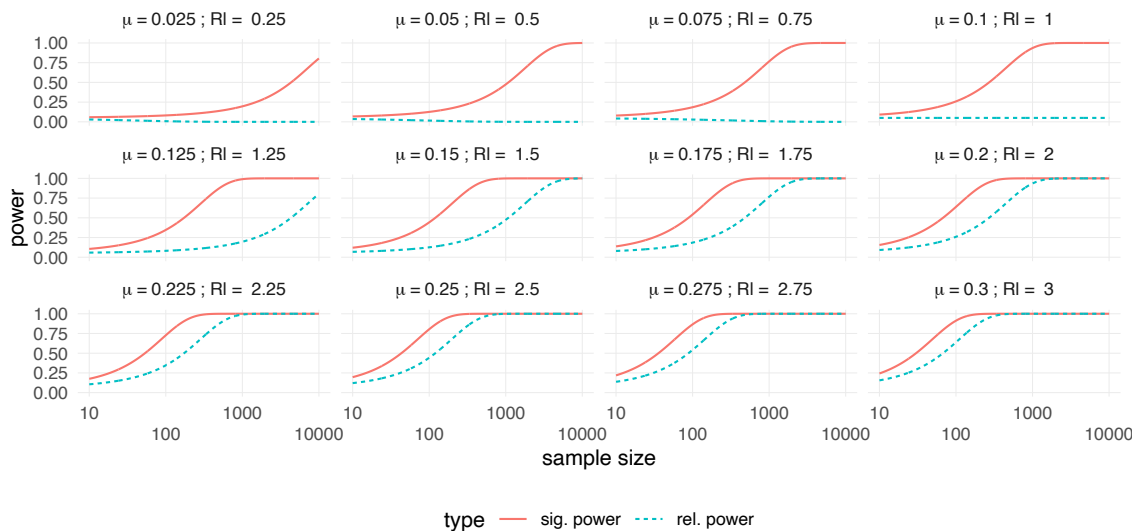


Figure 2.2: Rl.Power and Sig.Power over sample size.

As the LOR and the SMD are (approximately) normal distributed, so are their effect differences. We can thus compute valid (Wald) CIs. If the standard error (se) of both effect estimates are known we can pool the se 's to compute $se(\hat{\Delta})$ as:

$$se(\hat{\Delta}) = \sqrt{se(\hat{\theta}_1)^2 + se(\hat{\theta}_2)^2}$$

And use $se(\hat{\Delta})$ to construct a symmetric CI around Δ : From $\hat{\Delta} - z_\gamma \cdot se(\hat{\Delta})$ to $\hat{\Delta} + z_\gamma \cdot se(\hat{\Delta})$.

Anderson and Maxwell (2016) argue that evidence *for* equivalence of effect sizes should be based on a 90% CI of Δ as it conceptually corresponds to a TOST while evidence for a difference in effect size should consider a 95% CI. For this study, precision of estimates was fixed to 95% CIs. As effects of original studies have often been shown to be larger than replication effects, reported Rl in this study was based on the upper Relevance threshold. $Rl > 1$ would thus indicate that the original effect was relevantly larger than the replication effect.

Success classification

The scheme is very similar to the direct replication except we have now relevance thresholds in both directions. A successful replication with respect to effect magnitude would show equivalent effect sizes where the confidence interval of the differences is fully contained between the relevance thresholds. Because the CI of the difference in effects accounts for uncertainty of both the original study and the replication attempt the range of the confidence interval is generally larger than either of the two. As a relevance threshold of $\zeta = 0.1$ leads to the range $[-\zeta, \zeta] = [-0.1, 0.1]$, the range of the CI around the difference would have to be smaller than 0.2 to be fully contained within the relevance thresholds *if* the point estimates were to be exactly equal. Figure 2.3 shows the relation between the ranges of CIs of the original and the replication attempt, and the range of the CI of the difference under the assumption of equal point estimates (the best case scenario).

It is not immediately obvious what relevance threshold to chose as a default. Sticking

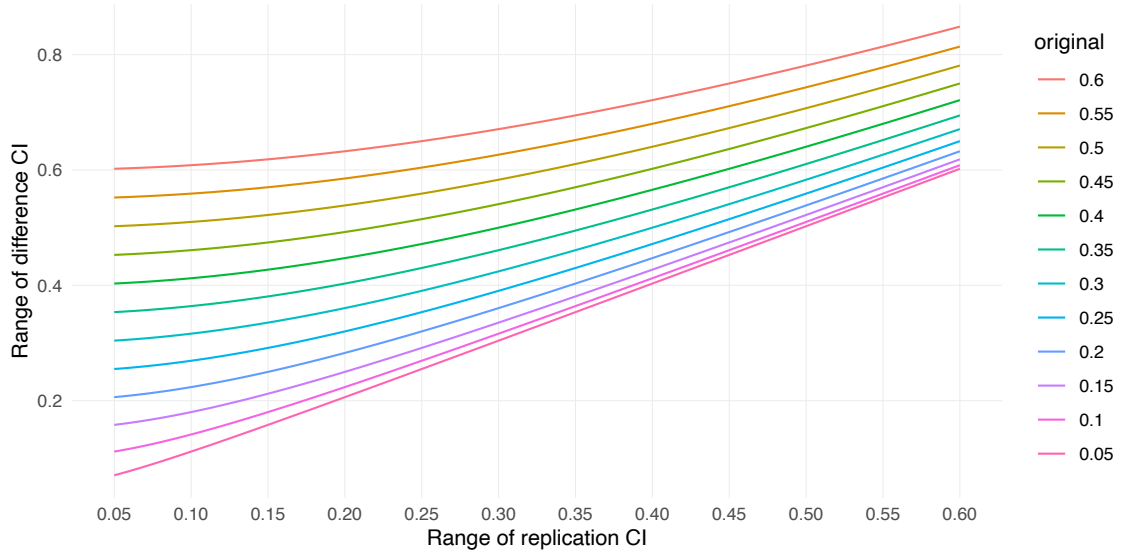


Figure 2.3: Best case of ranges of confidence intervals for effect differences given the CI ranges of the original and replication effects.

to the 10% threshold is possible but potentially too conservative. [Lakens \(2017\)](#) suggests for TOST procedures (until a better approach is found) to set the threshold such that we would have sufficient power to detect an effect of that size. As our studies were not powered for equivalence testing this might result in a very large thresholds and thus not capture the concept of Relevance well. For the sake of this study the threshold was set to $\zeta = 0.1$

2.4 Heterogeneity

A MEMo was fitted using the `lme4` package ([Bates et al., 2015](#)) to estimate the overall (meta-analytic) effect and its relevance. To determine whether the estimated effect of the original study was an anomaly, the MEMo estimated effect heterogeneity based only on the replication attempts. If the original study were an anomaly including it would unjustly inflate effect heterogeneity. Adherence to standardized study protocols ensured a conservative (best case) heterogeneity estimate for strict replications. Therefore, for the studies replicated in the Many Labs 5 project only the revised protocol attempts were included in the MEMo.

Heterogeneity of effect size was captured via the Intraclass Correlation Coefficient (*ICC*). The ICC relates the between group variance to the total variance:

$$ICC = \frac{\sigma_{\theta}^2}{\sigma_{total}^2}$$

The grouping variables in our study were the replication attempts (`location`). ICC was computed using the `specr` package ([Masur and Scharkow, 2019](#)). Within the logistic regression residual variance was set to $\sigma^2 = \pi^2/3$ ([Moineddin et al., 2007](#)).

Typically, researchers test whether there is significant heterogeneity. Instead, this study applied the concept of Relevance to the ICC. A relevance threshold of $\zeta = 0.1$ was chosen. Thus, a between group variance of 10% ($ICC = 0.1$) was considered relevant.

For the SMD and the logit, a prediction interval (PI) around the overall fixed effect was created following [Borenstein \(2009, pp. 129\)](#).

$$PI = \hat{\theta} \pm t_{df}^{\alpha} \sqrt{\sigma_{\hat{\theta}}^2 + se(\hat{\theta})^2},$$

Where $\hat{\theta}$ is the estimated effect, $\sigma_{\hat{\theta}}^2$ is the effect variance between the groups and $se(\hat{\theta})^2$ is the se of our estimate. [Borenstein \(2009\)](#) suggests degrees of freedom to be $df = k - 2$ where k is the number of replication attempts. This way, the PI is adjusted for heterogeneity and estimate uncertainty. The effect of a new replication attempt should then fall within the PI. If the replication attempts captured the same situation as the original study the originally estimated effect should fall within the PI. An original effect outside of the PI suggests that the original effect may be an anomaly (an approach similar to that of [Mathur and VanderWeele, 2020](#)).

Chapter 3

Data

3.1 Replication projects

Data from the Many Labs replication projects 1 and 5 was used for this re-analysis. In the Many Labs 1 project each of the 13 studies attempted to be replicated in every one of the 36 locations (Klein et al., 2014). This gave ample power to estimate effect heterogeneity of each effect as all attempts were conducted according to just one study protocol.

The Many Labs 5 project primarily investigated whether failed replications may be due to a suboptimal study protocol and whether conducting a registered report may increase replicability (Ebersole et al., 2020). Thus, for each study two study protocols were developed and carried out: The *replication protocol* was based on a previous replication project which had not been approved by independent peer review nor the authors of the original study. The *revised protocol* implemented improvements on the replication protocol as suggested by the authors of the original study. Having the authors of the original paper revise the study protocol ensured that the studies were conducted in the spirit of the original studies generally making them more direct replications. For the studies of the Many Labs 5 project the analysis focuses on the revised protocol data although the replication protocol data.

In this paper, *study* refers to one many labs replication study, e.g. the replication of the Schwarz et al. (1985) study conducted by Klein et al. (2014). The Original study refers to the original experiment (e.g. Schwarz et al., 1985). A replication attempt refers to one replication attempted by a specific research group at one particular location. A graphic presentation of the labeling is found in figure 3.1. Throughout the paper I will refer to variables in the data using the `code` font.

3.2 Studies

3.2.1 Albarracín et al. (2008) (alb5)

The experiment 5 of Albarracín et al. (2008) is a classic two group experimental setup. The research hypothesis was whether priming a person with *action* vs priming with *in-action* (the experimental `Condition`) would increase the number of correct solutions on a cognitive test featuring 21 questions assessing verbal ability and quantitative ability (`SATTotals`). Albarracín et al. (2008) found that participants primed with *action* achieved

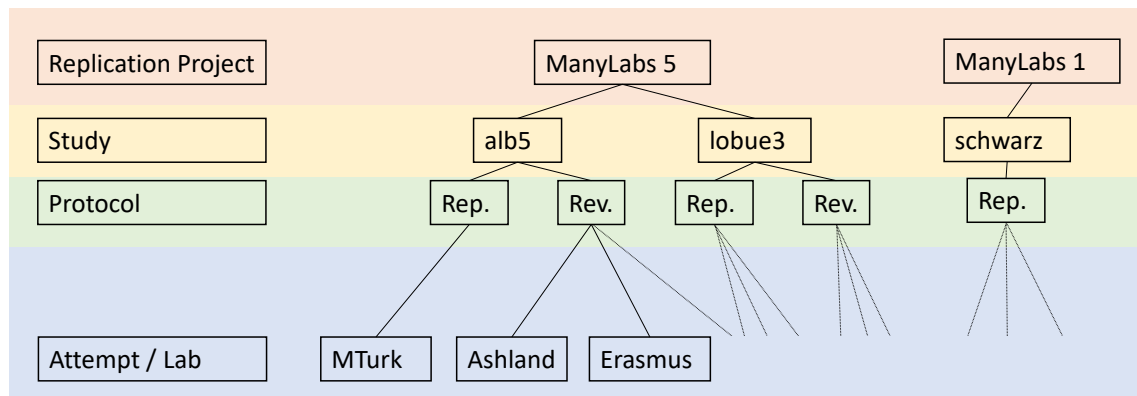


Figure 3.1: Notation and structure of the data.

a significantly higher number of correct results, $F(1, 34) = 5.68$, $p = 0.02$.

The main difference between the two study protocols was that the revised protocol was *not* conducted online but in person, just like the original study. For more differences, see the [osf page](#). This study was replicated within the Many Labs 5 project in eight locations using the revised protocol (with sample sizes N ranging from 81 to 174), and online through Amazon Mechanical Turk (*MTurk*) using the replication protocol [$N_{MTurk} = 580$; [Ebersole et al. \(2020\)](#)]. We used this study to exemplify the Relevance procedure for two independent samples, estimating the SMD.

3.2.2 Schwarz et al. (1985) (schwarz)

The study conducted by [Schwarz et al. \(1985\)](#) investigated the effect of the design of questionnaire-scales (`scalesgroup`) on the information reported by participants (un-intuitively labeled `scales`). Specifically, they randomly assigned $N = 132$ participants either into the low category or the high category group. Participants were asked to rate how many hours of tv they watched per day on a questionnaire. The answer options varied depending on the assigned group. In the low category group six answer categories ranging from *less than 0.5 hours* to *2.5 hours or more* were possible. In the high category group the six answer categories ranged from *less than 2.5 hours* to *4.5 hours or more*. They found that in the high category group significantly more people reported watching more than 2.5 hours of tv per day compared participants in the the low category group, $\chi^2(1) = 7.7$, $p < 0.005$.

The Many Labs 1 project conducted replication attempts in 13 locations and sample sizes ranged from $N = 71$ to $N = 1261$, totaling $N = 5899$ [Klein et al. \(2014\)](#). We used the schwarz study to illustrate LOR.

3.2.3 LoBue and DeLoache (2008) (lobue3)

Experiment three of the original study by [LoBue and DeLoache \(2008\)](#) was about attention to fear-relevant stimuli based on a sample of $N = 48$ participants, half of which were children. They measured reaction time (`RT.correct`) to fear relevant stimuli vs non fear relevant stimuli (`target_stimulus`: pictures of snakes vs. pictures of frogs or caterpillars) in children aged 3 and one of their parents (`child_parent`). Previous experience with snakes (`snake_experience`) was collected as a control variable. [LoBue and DeLoache \(2008\)](#) did find significant main effects of age, target stimuli, and its interaction, but no

effect for snake experience. [Lazarevic et al. \(2019\)](#) summarized the effect of the target stimuli of the original study as $R^2 = .23$, $CI[0.07; 0.49]$.

This study was replicated within the Many Labs 5 project ([Lazarevic et al., 2019](#)). Here, both study protocols were conducted in each of the four locations. In the revised protocol sample sizes ranged from $N = 53$ to $N = 71$ with a total of $N = 246$ participants. The replication protocol used pictures of caterpillars as non fear relevant stimuli and allowed children up to five years to take part while the revised protocol used pictures of frogs as non fear relevant stimuli and restricted the age of the children to three years of age like the original study. A full list of differences is available on the replication [OSF page](#). This study was re-analyzed using the adjusted coefficient of determination R^2_{adj} .

Chapter 4

Results

4.1 Alb5: SMD

Enough information was provided by the original paper to reconstruct the standardized and the unstandardised effect (Albarracín et al., 2008).

A total of eight replication attempts were conducted based on the revised protocol and one on the replication protocol. In figure 4.1 A we see that while the original study produced a relevant effect size with $Rls > 1$, none of the replication attempts produced a relevant effect. In fact, only three of the eight attempts had $Rle > 1$ ($Rle_{Ghent} = 1.17$, $Rle_{Illinois} = 2.53$, and $Rle_{Wesleyan} = 1.56$) and only the SMD CI of the study conducted in Illinois did not include 0; $SMD_{Illinois} = 0.253$, $CI[0.031; 0.476]$, $p = 0.026$. With seven out of eight attempts being Amb. and the remaining Amb.Sig there is very little evidence against the null hypothesis that there is no relevant effect. See table A.1 for all estimates.

The studies were sufficiently powered to find both significance ($Sig.Power > 0.9$) and Relevance ($Re.Power > 0.8$) given the hypothesized effect size of the original study.

The combined evidence of all studies (All) using the revision protocol paints a similar picture: Typical for the very large total sample ($N = 884$) the increased precision lead to an Amb.Sig. result. Although $Rls < 1$, Rle is just over 1, as the mixed effects SMD is just larger than the Relevance threshold ($\zeta = 0.1$).

4.1.1 Effect magnitudes

As expected, the confidence intervals around the effect size differences are rather large, see figure 4.1 B. Only the replication attempt at Queens differed significantly (but not relevantly, Amb.Sig) from the original effect size estimate. This was the attempt with the only negative effect size estimate $Rl_{Queens} = -.606$ $CI[-11.39; 2.6]$. With all other attempts being Amb. we have no evidence suggesting the effects to be equivalent to the original effect size. Although the difference of each attempt lacked precision to postulate a relevant difference the fact that seven of the eight differences had $Rle > 1$ does suggest that original study may have been an anomaly. All results can be found in table A.2.

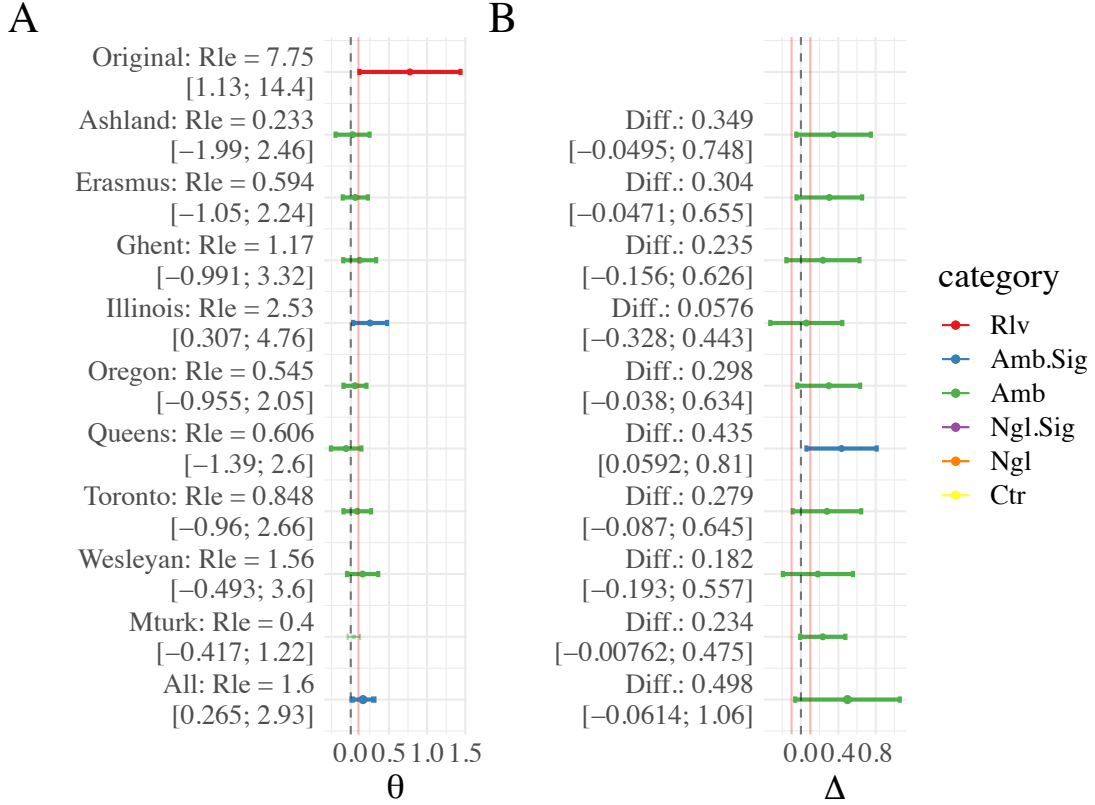


Figure 4.1: A: SMD estimates with CI and relevance threshold $\zeta = 0.1$. B: Difference between original effect size and each replication attempt.

4.1.2 Heterogeneity:

A mixed effects model with random intercept and slope without an intercept-slope covariance was fitted to estimate effect heterogeneity. While there was a relevant intercept variance ($Rl_{ICC_{Intercept}} = 1.8$) the ICC for the slope and its relevance was estimated to be 0. This is not surprising considering the effect estimates of the replication attempts in figure 4.1 A were all relatively close together with overlapping confidence intervals. Hence, the 95% PI around the fixed effects estimate was only slightly wider than the CI. We would expect the effect of a further (direct) replication to be within this PI. The original effect was not even close to the bounds of the MEMo PI which further suggested the original effect to be an anomaly.

4.1.3 Comparison to ManyLabs

It is difficult to compare the results of our re-analysis to the analysis by Ebersole et al. (2020) as the overarching paper did not focus on individual replication studies and the study-specific paper was not yet available. For the overarching paper the effects of all Many Labs 5 studies were transformed to *pearsons r*. While Ebersole et al. (2020) conceded that the replication effects were much smaller than the original effect, the alb5 study was still considered successful as the meta-analytic effect was significant. Our re-analysis concluded that if there were an effect, we could find no evidence suggesting it was of a relevant magnitude.

4.2 Schwarz: Log Odds Ratio

The LOR of the original study could be calculated from the table published in the original manuscript (Schwarz et al., 1985).

Figure 4.2 A shows the logit and the corresponding CI for each location. Estimated Sig.power for the studies based on 1000 simulations ranged between 59.4% and 100% and Rl.power between 51.6% and 100%. The Relevance estimate of the original study was clearly relevant, $Rle = 11.3$, $CI[3.06; 19.7]$.

10 effect estimates showed extremely wide CIs which had to be cut off in the plot. For some of them, the point estimate is therefore not visible but can be inferred from the corresponding Rle in the y-axis label. All cases where CIs were inflated were due to very few participants in the *low* category reporting to be watching 2.5h of tv or more per day. This naturally lead to very high standard errors.

15 studies produced relevant parameter estimates ($Rls > 1$, see figure 4.2). 18 studies (including the 10 studies with extremely large CIs) produced ambivalent results, and three produced ambivalent results with a significant effect (Amb.Sig). Only one of the 36 replication studies resulted in an effect estimate in the opposite direction, this effect was neither relevant nor significant, $Rle = -9.7$, $CI[-32.3; 14.2]$. Results of the directional replication can be found in table A.3.

4.2.1 Differences in effect

Results of the effect differences can be found in table A.4 and are visualized in figure 4.2 B. Although some estimates are relatively close to the estimate of the original effect size, the precision of the estimates are generally too low to determine whether or not the differences were relevant or negligible (Amb). However, most point estimates were larger than the original effect size which indicated that the original estimate was likely not exaggerated.

4.2.2 Heterogeneity:

Although the ICC of the intercept was relevant ($ICC_{intercept}$: $Rle = 1.08$) there was no relevant heterogeneity for the group ($ICC_{scalesgroup}$: $Rle = 0.21$). The PI around the fixed effects estimate did include the effect of the original study suggesting the original study not to be an anomaly.

4.2.3 Comparison to ManyLabs

While Schwarz et al. (1985) reported the effect as χ^2 , the replication study by Klein et al. (2014) also computed χ^2 then transformed it and reported cohen's d . As the study was framed to be interested in effect magnitudes applying logistic regression seemed a more natural solution (see also Herger and Rogai, 2018). Klein et al. (2014) only discussed the meta-analytic effect and not each individual replication attempt. They found that the replication overall had been successful as they found effects which were similar in magnitude than the original effect. Establishing equivalence between the original and the replication study was not attempted. The conclusions drawn overall were very similar to those of the present study: Direct replication were largely successful and effect magnitudes were similar.

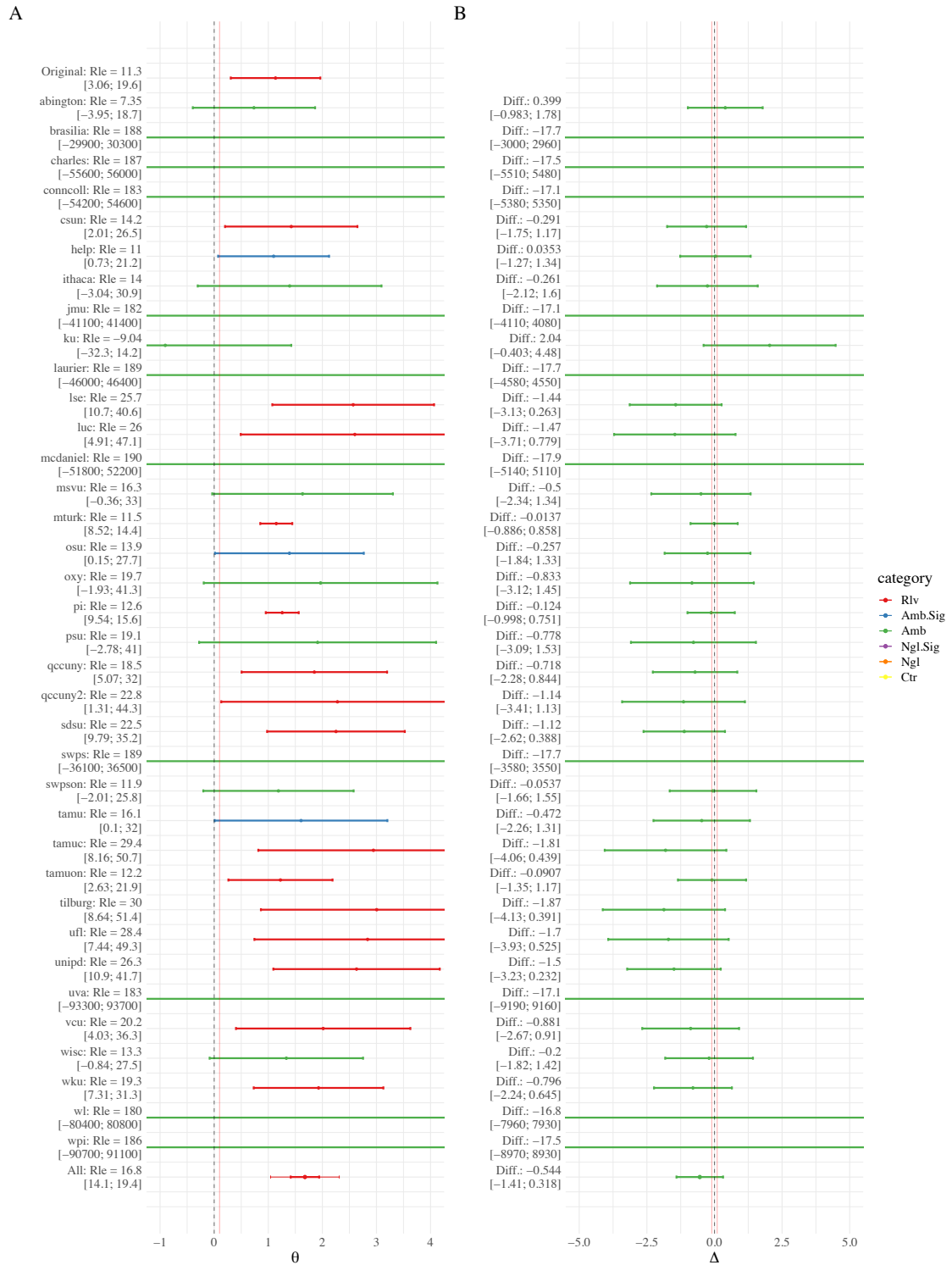


Figure 4.2: A: Logit estimates with CI and relevance threshold $\zeta = 0.1$. B: Difference between original effect size and each replication attempt.

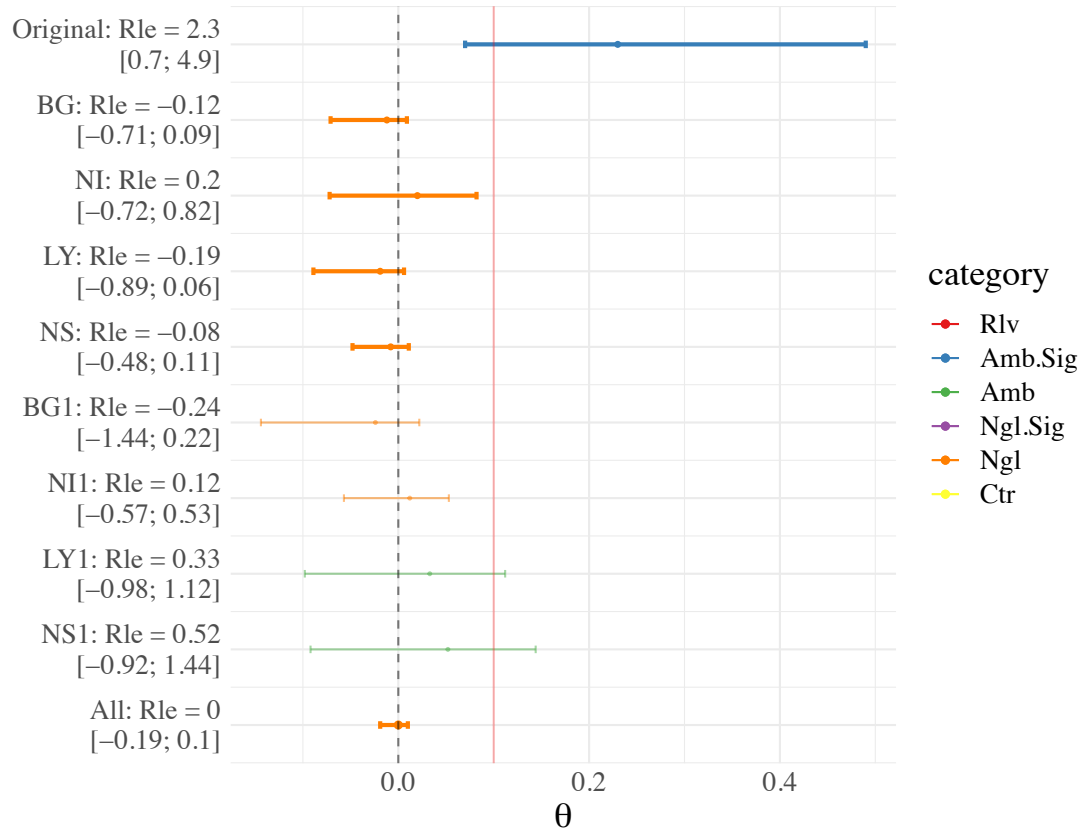


Figure 4.3: R^2 estimates with CI and relevance threshold for the lobue3 study, $\zeta = 0.1$.

4.3 Lobue3: R^2_{adj}

The measure of the original study was not based on exactly the same model which was fitted here. Computed differences between the replication attempts and the original study were therefore obsolete. A directional effect was still expected in the replication studies.

The original study effect was clearly significant but *not* relevant $Rle = 2.3$ $CI[0.7; 4.9]$. Only four replication attempts were conducted for each protocol. In the revised protocol all attempts showed a negligible effect of `target_stimulus` and its interactions on improving the predictive quality of the model (see figure 4.3). This can be viewed as clear evidence against the effect found in the original study. Two of the four replication protocol studies also showed Ngl results while the other two were Amb with $Rle > 1$. Heterogeneity was not estimated. The increase of R^2_{adj} in the MEMo was also negligible and estimated to be practically 0.

This study provided clear evidence against the efficacy of the `target_stimulus` both in the individual replication attempts and in the MEMo. Results of the lobue study are shown in table A.5.

4.3.1 Comparison to ManyLabs

Lazarevic et al. (2019) used data of both protocols (without using protocol as a covariate) in their replication meta-analysis. Only later they investigated whether protocol moderated effects; which it did. Systematically introducing variation in procedures and

subsequently not including it to test the replication hypothesis potentially reduces power to detect an effect and should be avoided. While [Lazarevic et al. \(2019\)](#) found no evidence of a significant effect of `target_stimulus` on reaction time, we found evidence that there is no (or only a negligible) effect, an epistemologically more potent conclusion.

Chapter 5

Conclusion

Summary of results

The three examples provided showed mixed results. For the alb5 study, a relevant directional effect could not be established. Although it appears that there was no relevant effect, precision was not high enough to establish the effect to be negligible at the chosen threshold. Because the precision of the original study was quite low a relevant difference between the original and the replication attempts could not be established. However, meta-analytic results showed the original effect to be much larger than what we would expect.

About half the attempts to replicate the schwarz study successfully replicated the relevant directional effect of the logit, providing ample evidence in favour of the original study. The other half were mostly ambivalent due to either inflated standard errors or general lack of precision. As no relevant effect heterogeneity was present effects were more or less similar. Again, precision was too small to make a conclusive statement on the difference of effects but since the original effect was covered in the MEMo PI it was likely no anomaly.

The lobue3 study was a bit different. While the point estimate of the original R^2 was relevant it was not conclusively so. As all effects of the (revised) replication attempts proved negligible, we could conclude that the additional variables contained no relevant predictive information. Due to the conceptual difference between the original study effect and the replication attempts, effect magnitudes were not formally compared.

Sometimes, comparing the effect size of the replication attempts to the effect size of the original study is not very constructive for two reasons: First, the precision is limited by the original study. For imprecise original estimates, this makes establishing equivalence practically impossible and determining a relevant difference between the studies unlikely if there is no substantial difference. Second, if the original study did not use *exactly* the same procedure and methods, differing effect sizes would be no surprise. The more variables differ between studies the higher we would expect effect heterogeneity to be (Yarkoni, 2019). If multiple replication attempts were made, effect heterogeneity and population level effects should be estimated to better understand an effect, its generalizability, and its limits. The prediction interval around that estimate can then be used to informally evaluate whether an original effect was indeed an anomaly.

A word on power

Estimating power based on the original effect size may be problematic if the original effect was an anomaly. Especially a small sample size and a somewhat surprising effect should deter replicators from relying on the original effect estimate to get an appropriate sample size. In our case the effect sizes of the original studies, as basis for the power calculations, were relatively large. A more conservative power estimate could rely for example on the lower bound of the CI of the original effect.

Further, a useful replication study should be powered to both detect a relevant effect but also to judge an effect as conclusively negligible. $1 - \alpha$ (95%) CIs were used for both decisions. Using $1 - 2 * \alpha$ (90%) CIs for categorizations on equivalence (or negligibility) as suggested by [Lakens \(2017\)](#) would produce less ambivalent results. For more conclusive results, power should also be calculated for negligible outcomes. For example, one could aim for a power of 80% to conclude an effect to be negligible given that the (standardized) effect size is at most $\zeta/2$ ($Rl = 0.5$).

On the other hand, the alb5 replication attempts included a total of 884 participants and provided merely ambivalent (or Amb.sig) results even on a meta-analytic level without any effect heterogeneity present. If with 884 participants a relevant effect could not be established I am inclined to call it negligible nonetheless. This suggests that the threshold for negligibility was very much too conservative. It is not feasible — nor does it seem desirable — to conduct such replication studies (or even original studies) using so many resources.

A less conservative threshold for equivalence could be chosen. If there is no theory driven or empirical equivalence threshold justification one could base it on the resulting sample size of the directional relevance power calculation and a hypothesized effect of $Rl = 0.5$. One could thus define an equivalence threshold as the 80% quantile of the simulated distribution of an effect with $Rl = 0.5$ and a prespecified sample size. We could then expect 80% of replication attempts to be classified as negligible if the true effect would be $Rl = 0.5$ or lower. Unfortunately, this may lead to an unacceptably large threshold. Although there is no one correct way to specify a threshold [Lakens et al. \(2018\)](#) offer three useful justifications for what they call *smallest effect size of interest*.

Limitations and benefits

By requiring specification of a Relevance threshold, an initial effort from researchers is required, which adds another degree of freedom to an analysis [Stahel \(2021b\)](#). Further, the choice of a threshold affects the Rl parameter in a multiplicative way. In practice, this is somewhat countered by the introduction of sensible default values, where deviations thereof should be justified to some extent. While for equivalence testing specifying a threshold in some way has long been accepted, asserting the use of thresholds in a place where NHST is still widely accepted would come close to a paradigm shift. In addition, Relevance is generally more conservative and resource intensive than NHST; A costly demand to make.

However, the costs are justified: Relevance, as a single parameter, shifts the focus away from precision and represents information both for *and* against the null hypothesis. It also conveys the effect size on an intuitive scale directly related to a minimally interesting effect size. By offering (at least) six different categorizations of estimates of a parameter it provides a much more nuanced assessment of evidence than NHST. All this while being

easily understood (both conceptually and mathematically) and without leaving the safe haven of frequentist theory.

These benefits were demonstrated in the present study: The Relevance parameter was able to communicate the successful directional replication of schwarz, and provided evidence against the effect found in the lobue study. All without overreaching the limits of its conceptual validity, as shown by the many ambivalent results (mainly for comparing effect sizes).

Outlook

To facilitate widespread adoption of the measure there is still more work to be done. Guidelines or rules of thumb for the choice of relevance and equivalence thresholds need to be established. It is not yet clear whether the threshold should vary between evidence for and against the null hypothesis. The ability to categorize effects to be negligible is especially important for replication studies. A thorough and empirical comparison between various alternatives of the NHST procedure (e.g. second generation p-values, bayesian methods and more) would be very useful.

While the `replication` package is very user friendly, users do have to engage with the package and its functions directly. Some packages (e.g. `lmerTest` [Kuznetsova et al., 2017](#)) appropriate commonly used function calls when loaded and autonomously supplement their output. This would be a subtle way to ease access and nudge users towards familiarity with the Relevance parameter without having to convince individual package authors to approve and implement the measures first. Another way forward would be to cooperate with influential researchers endorsing the measure and teaching students.

Of course, Relevance within replicability is not limited to replication studies. Instead, applying relevance through re-analysis of original studies can already give an indication whether spending resources on a replication of that particular study is worth the effort. If a (large sample) study presents a significant effect which can be shown to be either negligible in size or has $Rls < 1$ or even $Rle < 1$ I would not have high hopes of replicating such an effect. Future studies should systematically investigate whether such constellations are common and whether an originally relevant effect size is generally more likely to replicate.

Bibliography

- Ahn, S., Myers, N. D., and Jin, Y. (2012). Use of the estimated intraclass correlation for correcting differences in effect size by level. *Behavior Research Methods*, 44(2):490–502.
- Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., Brown, R. D., Earl, A., and Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology*, 95(3):510–523.
- Anderson, S. F. and Maxwell, S. E. (2016). There’s more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1):1–12.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10.
- Bland, J. M. and Altman, D. G. (2000). The odds ratio. *BMJ : British Medical Journal*, 320(7247):1468.
- Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., and Dupont, W. D. (2019). An Introduction to Second-Generation p-Values. *The American Statistician*, 73(sup1):157–167.
- Bonett, D. G. (2021). Design and Analysis of Replication Studies. *Organizational Research Methods*, 24(3):513–529.
- Borenstein, M. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Chichester, West Sussex, U.K. ;.
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., and van ’t Veer, A. (2014). The Replication

- Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50:217–224.
- Center for Open Science (2016). Templates of OSF Registration Forms.
- Center for Open Science (2021). Registered Reports. <https://www.cos.io/initiatives/registered-reports>.
- Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).
- Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12):171085.
- Cumming, G. (2008). Replication and p Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, 3(4):286–300.
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrichetto, L., Arnal, J. D., Arrow, H., Babincak, P., Bakos, B. E., Baník, G., Baskin, E., Belopavlović, R., Bernstein, M. H., Bialek, M., Bloxson, N. G., Bodroža, B., Bonfiglio, D. B. V., Boucher, L., Brühlmann, F., Brumbaugh, C. C., Casini, E., Chen, Y., Chiorri, C., Chopik, W. J., Christ, O., Ciunci, A. M., Claypool, H. M., Coary, S., Čolić, M. V., Collins, W. M., Curran, P. G., Day, C. R., Dering, B., Dreber, A., Edlund, J. E., Falcão, F., Fedor, A., Feinberg, L., Ferguson, I. R., Ford, M., Frank, M. C., Fryberger, E., Garinther, A., Gawryluk, K., Ashbaugh, K., Giacomantonio, M., Giessner, S. R., Grahe, J. E., Guadagno, R. E., Hałasa, E., Hancock, P. J. B., Hilliard, R. A., Hüffmeier, J., Hughes, S., Idzikowska, K., Inzlicht, M., Jern, A., Jiménez-Leal, W., Johannesson, M., Joy-Gaba, J. A., Kauff, M., Kellier, D. J., Kessinger, G., Kidwell, M. C., Kimbrough, A. M., King, J. P. J., Kolb, V. S., Kołodziej, S., Kovacs, M., Krasuska, K., Kraus, S., Krueger, L. E., Kuchno, K., Lage, C. A., Langford, E. V., Levitan, C. A., de Lima, T. J. S., Lin, H., Lins, S., Loy, J. E., Manfredi, D., Markiewicz, Ł., Menon, M., Mercier, B., Metzger, M., Meyet, V., Millen, A. E., Miller, J. K., Montealegre, A., Moore, D. A., Muda, R., Nave, G., Nichols, A. L., Novak, S. A., Nunnally, C., Orlić, A., Palinkas, A., Panno, A., Parks, K. P., Pedović, I., Pękala, E., Penner, M. R., Pessers, S., Petrović, B., Pfeiffer, T., Pieńkosz, D., Preti, E., Purić, D., Ramos, T., Ravid, J., Razza, T. S., Rentzsch, K., Richetin, J., Rife, S. C., Rosa, A. D., Rudy, K. H., Salamon, J., Saunders, B., Sawicki, P., Schmidt, K., Schuepfer, K., Schultze, T., Schulz-Hardt, S., Schütz, A., Shabazian, A. N., Shubella, R. L., Siegel, A., Silva, R., Sioma, B., Skorb, L., de Souza, L. E. C., Steegen, S., Stein, L. A. R., Sternglanz, R. W., Stojilović, D., Storage, D., Sullivan, G. B., Szaszi, B., Szecsi, P., Szöke, O., Szuts, A., Thomae, M., Tidwell, N. D., Tocco, C., Torka, A.-K., Tuerlinckx, F., Vanpaemel, W., Vaughn, L. A., Vianello, M., Viganola, D., Vlachou, M., Walker, R. J., Weissgerber, S. C., Wichman, A. L., Wiggins, B. J., Wolf, D., Wood, M. J., Zealley, D., Žeželj, I., Zrubka, M., and 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.
- Fletcher, S. C. (2021). How (not) to measure replication. *European Journal for Philosophy of Science*, 11(2):57.
- Francis, G. and Thunell, E. (2020). Excess success in “Don’t count calorie labeling out:

- Calorie counts on the left side of menu items lead to lower calorie food choices". *Meta-Psychology*, 4.
- Gibson, E. W. (2021). The Role of p-Values in Judging the Strength of Evidence and Realistic Replication Expectations. *Statistics in Biopharmaceutical Research*, 13(1):6–18.
- Hayes, A. and Moller-Trane, R. (2019). *distributions3: Probability Distributions as S3 Objects*. R package version 0.1.1.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3):e1002106.
- Hedges, L. V. and Schauer, J. M. (2019). More Than One Replication Study Is Needed for Unambiguous Tests of Replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570.
- Held, L. (2019). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*, 6(3):181534.
- Herger, L. and Rogai, F. (2018). *Replicability of Psychological Studies: A New Analysis of the ManyLabs Data Set*. Seminar Work, ETH Zurich.
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124.
- Kenny, D. A. and Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, 24(5):578–589.
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3):196–217.
- Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Joy-Gaba, J., Legg, A. M., Levitan, C., Mancini, A., Manfredi, D., Miller, J. M., Nave, G., Redford, L., Schlitz, I., Schmidt, K., Skorinko, J., Storage, D., Swanson, T., van Swol, L., Vaughn, L. A., and Ratliff, K. (2019). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Ann Vaughn, L., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, 45(3):142–152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching,

- R., Rédei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E., Cheong, W., Cicero, D. C., Coen, S., Coleman, J. A., Collisson, B., Conway, M. A., Corker, K. S., Curran, P. G., Cushman, F., Dagona, Z. K., Dalgard, I., Dalla Rosa, A., Davis, W. E., de Bruijn, M., De Schutter, L., Devos, T., de Vries, M., Doğulu, C., Dozo, N., Dukes, K. N., Dunham, Y., Durrheim, K., Ebersole, C. R., Edlund, J. E., Eller, A., English, A. S., Finck, C., Frankowska, N., Freyre, M.-Á., Friedman, M., Galliani, E. M., Gandi, J. C., Ghoshal, T., Giessner, S. R., Gill, T., Gnamb, T., Gómez, Á., González, R., Graham, J., Grahe, J. E., Grahek, I., Green, E. G. T., Hai, K., Haigh, M., Haines, E. L., Hall, M. P., Heffernan, M. E., Hicks, J. A., Houdek, P., Huntsinger, J. R., Huynh, H. P., IJzerman, H., Inbar, Y., Innes-Ker, Å. H., Jiménez-Leal, W., John, M.-S., Joy-Gaba, J. A., Kamiloğlu, R. G., Kappes, H. B., Karabati, S., Karick, H., Keller, V. N., Kende, A., Kervyn, N., Knežević, G., Kovacs, C., Krueger, L. E., Kurapov, G., Kurtz, J., Lakens, D., Lazarević, L. B., Levitan, C. A., Lewis, N. A., Lins, S., Lipsey, N. P., Losee, J. E., Maassen, E., Maitner, A. T., Malingumu, W., Mallett, R. K., Marotta, S. A., Mededović, J., Mena-Pacheco, F., Milfont, T. L., Morris, W. L., Murphy, S. C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A. J., Neto, F., Lee Nichols, A., Ocampo, A., O'Donnell, S. L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Pérez-Sánchez, R., Petrović, B., Pilati, R., Pinter, B., Podesta, L., Pogge, G., Pollmann, M. M. H., Rutchick, A. M., Saavedra, P., Saeri, A. K., Salomon, E., Schmidt, K., Schönbrodt, F. D., Sekerdej, M. B., Sirlopú, D., Skorinko, J. L. M., Smith, M. A., Smith-Castro, V., Smolders, K. C. H. J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T. G., Stouten, J., Street, C. N. H., Sundfelt, O. K., Szeto, S., Szumowska, E., Tang, A. C. W., Tanzer, N., Tear, M. J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J. M., Ujhelyi, A., van Aert, R. C. M., van Assen, M. A. L. M., van der Hulst, M., van Lange, P. A. M., van 't Veer, A. E., Vásquez-Echeverría, A., Ann Vaughn, L., Vázquez, A., Vega, L. D., Verniers, C., Verschoor, M., Voermans, I. P. J., Vranka, M. A., Welch, C., Wichman, A. L., Williams, L. A., Wood, M., Woodzicka, J. A., Wronska, M. K., Young, L., Zelenski, J. M., Zhijia, Z., and Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.
- Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4):355–362.
- Lakens, D. (2020). The 20% Statistician: Review of "The Generalizability Crisis" by Tal Yarkoni.
- Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.
- Lazarevic, L., Ebersole, C. R., Nosek, B. A., Zezelj, I., and Purić, D. (2019). Many Labs 5: Registered Replication Report of LoBue & DeLoache (2018).
- Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E.-J., and van Ravenzwaaij, D. (2020). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor.

- LoBue, V. and DeLoache, J. S. (2008). Detecting the Snake in the Grass: Attention to Fear-Relevant Stimuli by Adults and Young Children. *Psychological Science*, 19(3):284–289.
- Martin, G. N. and Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology*, 0.
- Masur, P. K. and Scharnow, M. (2019). `specr`: Statistical functions for conducting specification curve analyses (version 0.2.1).
- Mathur, M. B. and VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1145–1166.
- Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70(6):487–498.
- Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, 34(2):103–115.
- Moineddin, R., Matheson, F. I., and Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 7(1):34.
- Nosek, B. A. and Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, 45(3):137–141.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.
- Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985). Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49(3):388–395.
- Stahel, W. (2021a). Replicability: Terminology, Measuring Success, Strategy. *Unpublished*.
- Stahel, W. A. (2021b). New relevance and significance measures to replace p-values. *PLOS ONE*, 16(6):e0252991.
- Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, 41(2):91–103.
- Thoma, S. P. (2021). *ReplicationRelevance: Calculates Relevance for Many Labs Replication Projects*. R package version 0.2.0.
- Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.
- Wikipedia (2021a). ClinicalTrials.gov. *Wikipedia*.
- Wikipedia (2021b). Coefficient of determination. *Wikipedia*.
- Williams, C. R. (2019). How redefining statistical significance can worsen the replication crisis. *Economics Letters*, 181:65–69.

Yarkoni, T. (2019). The Generalizability Crisis.

Zhang, D. (2021a). Coefficients of Determination for Mixed-Effects Models.
arXiv:2007.08675 [stat].

Zhang, D. (2021b). *rsq: R-Squared and Related Measures*. R package version 2.2.

Appendix A

Tables

A.1 Alb5

	Location	Mean Difference	SMD	Rle	N	P.power	RL.power	Class.
	original	2.05 [0.3; 3.8]	0.775 [0.113;1.437]	7.749 [1.132; 14.365]	36			Rlv
rev.	Ashland	0.145 [-1.245; 1.536]	0.023 [-0.199;0.246]	0.233 [-1.993; 2.458]	81	0.903	0.813	Amb
	Erasmus	0.358 [-0.632; 1.348]	0.059 [-0.105;0.224]	0.594 [-1.048; 2.235]	146	0.990	0.981	Amb
	Ghent	0.745 [-0.633; 2.124]	0.117 [-0.099;0.332]	1.166 [-0.991; 3.323]	86	0.946	0.884	Amb
	Illinois	1.714 [0.208; 3.221]	0.253 [0.031;0.476]	2.532 [0.307; 4.757]	81	0.916	0.838	Amb.Sig
	Oregon	0.34 [-0.596; 1.276]	0.055 [-0.096;0.205]	0.545 [-0.955; 2.046]	174	0.999	0.992	Amb
	Queens	-0.386 [-1.656; 0.885]	-0.061 [-0.26;0.139]	0.606 [-1.389; 2.6]	100	0.959	0.907	Amb
	Toronto	0.517 [-0.586; 1.62]	0.085 [-0.096;0.266]	0.848 [-0.96; 2.655]	121	0.990	0.965	Amb
	Wesleyan	1.012 [-0.32; 2.344]	0.156 [-0.049;0.36]	1.556 [-0.493; 3.604]	95	0.974	0.914	Amb
rep.	Mturk	0.306 [-0.318; 0.93]	0.04 [-0.042;0.122]	0.4 [-0.417; 1.216]	580	1.000	1.000	Amb
MEMo	All	0.501 [0.083; 0.917]	0.16 [0.027;0.293]	1.602 [0.265; 2.932]	884			Amb.Sig

Table A.1: Effect estimates and relevance of the alb5 replication analysis.

Location	Difference	Std. Difference	Rle	Class.
Ashland	1.905 [-0.27; 4.08]	0.349 [-0.05;0.748]	3.491 [-0.495; 7.477]	Amb
Erasmus	1.692 [-0.262; 3.646]	0.304 [-0.047;0.655]	3.038 [-0.471; 6.547]	Amb
Ghent	1.305 [-0.863; 3.473]	0.235 [-0.156;0.626]	2.353 [-1.556; 6.262]	Amb
Illinois	0.336 [-1.912; 2.584]	0.058 [-0.328;0.443]	0.576 [-3.277; 4.429]	Amb
Oregon	1.71 [-0.218; 3.638]	0.298 [-0.038;0.634]	2.978 [-0.38; 6.337]	Amb
Queens	2.436 [0.332; 4.54]	0.435 [0.059;0.81]	4.346 [0.592; 8.101]	Amb.Sig
Toronto	1.533 [-0.479; 3.545]	0.279 [-0.087;0.645]	2.789 [-0.87; 6.448]	Amb
Wesleyan	1.038 [-1.103; 3.179]	0.182 [-0.193;0.557]	1.819 [-1.933; 5.571]	Amb
Mturk	1.744 [-0.057; 3.545]	0.234 [-0.008;0.475]	2.339 [-0.076; 4.755]	Amb
All	1.549 [-0.191; 3.289]	0.498 [-0.061;1.058]	4.981 [-0.614; 10.576]	Amb

Table A.2: Alb5 effect magnitude comparison between replication attempts and the original study.

A.2 Schwarz

Location	Log OR	Rle	N	P.power	Rl.power	Class.
Original	1.134 [0.306; 1.962]	11.343 [3.064; 19.622]	132.000			Rlv
abington	0.735 [-0.395; 1.866]	7.35 [-3.95; 18.66]	79.000	0.674	0.606	Amb
brasilia	18.801 [-2994.86; 3032.461]	188.01 [-29948.6; 30324.61]	113.000	0.793	0.716	Amb
charles	18.679 [-5564.811; 5602.169]	186.79 [-55648.11; 56021.69]	78.000	0.645	0.562	Amb
conncoll	18.263 [-5422.285; 5458.812]	182.63 [-54222.85; 54588.12]	86.000	0.687	0.615	Amb
csun	1.425 [0.201; 2.648]	14.25 [2.01; 26.48]	88.000	0.714	0.625	Rlv
help	1.099 [0.073; 2.124]	10.99 [0.73; 21.24]	94.000	0.737	0.658	Amb.Sig
ithaca	1.395 [-0.304; 3.093]	13.95 [-3.04; 30.93]	76.000	0.637	0.537	Amb
jmu	18.227 [-4108.222; 4144.675]	182.27 [-41082.22; 41446.75]	163.000	0.934	0.876	Amb
ku	-0.904 [-3.234; 1.425]	-9.04 [-32.34; 14.25]	97.000	0.768	0.698	Amb
laurier	18.874 [-4599.473; 4637.222]	188.74 [-45994.73; 46372.22]	103.000	0.763	0.693	Amb
lse	2.57 [1.075; 4.065]	25.7 [10.75; 40.65]	260.000	0.994	0.971	Rlv
luc	2.6 [0.491; 4.71]	26 [4.91; 47.1]	126.000	0.859	0.786	Rlv
mcDaniel	18.986 [-5178.745; 5216.716]	189.86 [-51787.45; 52167.16]	87.000	0.690	0.597	Amb
msvu	1.634 [-0.036; 3.304]	16.34 [-0.36; 33.04]	78.000	0.658	0.588	Amb
mturk	1.148 [0.852; 1.444]	11.48 [8.52; 14.44]	983.000	1.000	1.000	Rlv
osu	1.391 [0.015; 2.766]	13.91 [0.15; 27.66]	99.000	0.768	0.687	Amb.Sig
oxy	1.967 [-0.193; 4.128]	19.67 [-1.93; 41.28]	101.000	0.786	0.712	Amb
pi	1.258 [0.954; 1.563]	12.58 [9.54; 15.63]	1261.000	1.000	1.000	Rlv
psu	1.912 [-0.278; 4.103]	19.12 [-2.78; 41.03]	90.000	0.722	0.646	Amb
qccuny	1.852 [0.507; 3.197]	18.52 [5.07; 31.97]	100.000	0.777	0.695	Rlv
qccuny2	2.279 [0.131; 4.428]	22.79 [1.31; 44.28]	83.000	0.700	0.624	Rlv
sdsu	2.252 [0.979; 3.524]	22.52 [9.79; 35.24]	154.000	0.919	0.863	Rlv
swps	18.873 [-3607.459; 3645.205]	188.73 [-36074.59; 36452.05]	71.000	0.656	0.552	Amb
swpson	1.188 [-0.201; 2.578]	11.88 [-2.01; 25.78]	127.000	0.876	0.805	Amb
tamu	1.606 [0.01; 3.202]	16.06 [0.1; 32.02]	156.000	0.928	0.880	Amb.Sig
tamuc	2.944 [0.816; 5.073]	29.44 [8.16; 50.73]	78.000	0.662	0.575	Rlv
tamuon	1.225 [0.263; 2.188]	12.25 [2.63; 21.88]	212.000	0.978	0.951	Rlv
tilburg	3.004 [0.864; 5.144]	30.04 [8.64; 51.44]	80.000	0.683	0.590	Rlv
ufl	2.836 [0.744; 4.928]	28.36 [7.44; 49.28]	118.000	0.834	0.787	Rlv
unipd	2.632 [1.094; 4.17]	26.32 [10.94; 41.7]	134.000	0.872	0.814	Rlv
uva	18.27 [-9330.205; 9366.746]	182.7 [-93302.05; 93667.46]	67.000	0.594	0.516	Amb
vcu	2.015 [0.403; 3.627]	20.15 [4.03; 36.27]	99.000	0.741	0.677	Rlv
wisc	1.334 [-0.084; 2.753]	13.34 [-0.84; 27.53]	91.000	0.744	0.657	Amb
wku	1.93 [0.731; 3.128]	19.3 [7.31; 31.28]	97.000	0.750	0.683	Rlv
wl	17.983 [-8039.404; 8075.369]	179.83 [-80394.04; 80753.69]	89.000	0.704	0.642	Amb
wpi	18.622 [-9068.46; 9105.703]	186.22 [-90684.6; 91057.03]	81.000	0.658	0.599	Amb
All	1.678 [1.415; 1.941]	16.78 [14.147; 19.412]	5899.000			Rlv

Table A.3: Effect estimates and relevance of the schwarz replication analysis.

Location	Log OR Difference	Rle	Class.
abington	0.399 [-0.983; 1.782]	3.993 [-9.834; 17.821]	Amb
brasilia	-17.667 [-2998; 2963]	-176.667 [-29984; 29631]	Amb
charles	-17.545 [-5512; 5477]	-175.447 [-55121.43; 54770.537]	Amb
conncoll	-17.129 [-5379; 5345]	-171.287 [-53793; 53450]	Amb
csun	-0.291 [-1.75; 1.169]	-2.907 [-17.503; 11.689]	Amb
help	0.035 [-1.267; 1.337]	0.353 [-12.668; 13.375]	Amb
ithaca	-0.261 [-2.121; 1.6]	-2.607 [-21.211; 15.998]	Amb
jmu	-17.093 [-4112; 4078]	-170.927 [-41125; 40783]	Amb
ku	2.038 [-0.403; 4.479]	20.383 [-4.026; 44.793]	Amb
laurier	-17.74 [-4580; 4545]	-177.397 [-45807; 45452]	Amb
lse	-1.436 [-3.134; 0.263]	-14.357 [-31.344; 2.631]	Amb
luc	-1.466 [-3.71; 0.779]	-14.657 [-37.102; 7.789]	Amb
mcdaniel	-17.852 [-5141; 5105]	-178.517 [-51415; 51058]	Amb
msvu	-0.5 [-2.336; 1.336]	-4.997 [-23.355; 13.362]	Amb
mturk	-0.014 [-0.886; 0.858]	-0.137 [-8.856; 8.583]	Amb
osu	-0.257 [-1.843; 1.33]	-2.567 [-18.434; 13.3]	Amb
oxy	-0.833 [-3.119; 1.454]	-8.327 [-31.192; 14.539]	Amb
pi	-0.124 [-0.998; 0.751]	-1.237 [-9.983; 7.51]	Amb
psu	-0.778 [-3.088; 1.533]	-7.777 [-30.88; 15.327]	Amb
qccuny	-0.718 [-2.279; 0.844]	-7.177 [-22.793; 8.439]	Amb
qccuny2	-1.145 [-3.415; 1.125]	-11.447 [-34.148; 11.254]	Amb
sdsu	-1.118 [-2.623; 0.388]	-11.177 [-26.23; 3.876]	Amb
swps	-17.739 [-3580; 3545]	-177.387 [-35804; 35450]	Amb
swpson	-0.054 [-1.655; 1.548]	-0.537 [-16.555; 15.481]	Amb
tamu	-0.472 [-2.255; 1.312]	-4.717 [-22.551; 13.118]	Amb
tamuc	-1.81 [-4.058; 0.439]	-18.097 [-40.579; 4.385]	Amb
tamuon	-0.091 [-1.351; 1.169]	-0.907 [-13.506; 11.693]	Amb
tilburg	-1.87 [-4.131; 0.391]	-18.697 [-41.306; 3.913]	Amb
ufl	-1.702 [-3.928; 0.525]	-17.017 [-39.28; 5.246]	Amb
unipd	-1.498 [-3.227; 0.232]	-14.977 [-32.274; 2.32]	Amb
uva	-17.136 [-9191; 9157]	-171.357 [-91916; 91573]	Amb
vcu	-0.881 [-2.671; 0.91]	-8.807 [-26.711; 9.097]	Amb
wisc	-0.2 [-1.822; 1.422]	-1.997 [-18.217; 14.224]	Amb
wku	-0.796 [-2.236; 0.645]	-7.957 [-22.359; 6.445]	Amb
wl	-16.849 [-7962; 7928]	-168.487 [-79621; 79284]	Amb
wpi	-17.488 [-8965; 8930]	-174.877 [-89653; 89304]	Amb
All	-0.544 [-1.405; 0.318]	-5.436 [-14.05; 3.178]	Amb

Table A.4: Schwarz effect magnitude comparison between replication attempts and the original study.

A.3 Lobue3

Location	Adj. R squared	Rle
Original	0.23 [0.07; 0.49]	2.3 [0.7; 4.9]
BG	-0.012 [-0.071; 0.009]	-0.12 [-0.71; 0.09]
NI	0.02 [-0.072; 0.082]	0.2 [-0.72; 0.82]
LY	-0.019 [-0.089; 0.006]	-0.19 [-0.89; 0.06]
NS	-0.008 [-0.048; 0.011]	-0.08 [-0.48; 0.11]
BG1	-0.024 [-0.144; 0.022]	-0.24 [-1.44; 0.22]
NI1	0.012 [-0.057; 0.053]	0.12 [-0.57; 0.53]
LY1	0.033 [-0.098; 0.112]	0.33 [-0.98; 1.12]
NS1	0.052 [-0.092; 0.144]	0.52 [-0.92; 1.44]
All	0 [-0.019; 0.01]	0 [-0.19; 0.1]

Table A.5: R squared estimates and relevance of the lobue3 replication analysis.

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

Estimating Relevance within Replication Studies

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):

First name(s):

Thoma

Stefan Pascal


With my signature I confirm that

- I have committed none of the forms of plagiarism described in the **Citation etiquette** information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics*.

Place, date:

Signature(s):

Bern, 24.8.2021



For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.