Training Materials

Introduction

This document contains a quick reminder of the aims and approach of the repliCATS project, followed by some training material on information that is likely to be useful to you when trying to understand and interpret research claims, and when you are assessing their replicability.

The repliCATS project

The University of Melbourne repliCATS team elicits expert judgements about the replicability of research claims in the Social and Behavioural Sciences through an online platform using the IDEA protocol. Judgements are aggregated into measures of reliability and the reasoning used is analysed. IDEA ("Investigate", "Discuss", "Estimate" and "Aggregate") has been found to improve judgements under uncertainty. More information about the repliCATS project is contained in the Plain Language Statement and on the repliCATS website.

The IDEA protocol

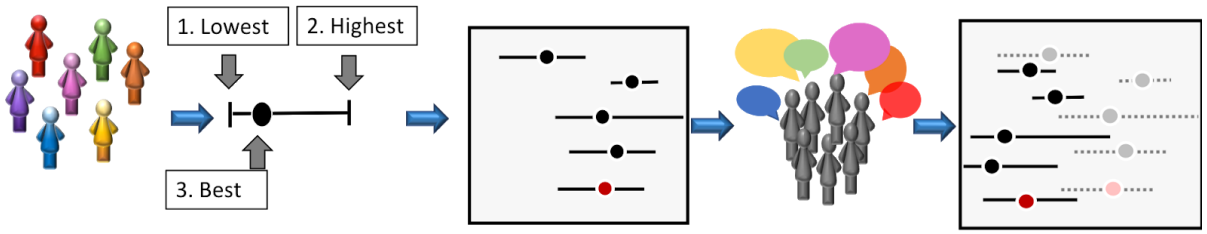
This protocol, developed at the University of Melbourne, has been found to improve judgements under uncertainty. IDEA stands for "Investigate", "Discuss", "Estimate" and "Aggregate", the four steps in the process of this elicitation.

For repliCATS, the IDEA protocol will involve participants:

1. Independently *Investigating* the claim, providing their personal judgement on the replicability of the claim, and commenting on their thinking.
2. Seeing the judgements of the rest of their team, the aggregated judgement and associated comments, and having a facilitated *Discussion* with the group. This phase can resolve uncertainties, and investigate evidence and thinking.
3. Providing a revised *Estimate* and describing how their thinking has changed.

The repliCATS team will use an *Aggregate* of the group judgements as the final assessment of the replicability of the research claim.

More information on the IDEA protocol can be found here.



1. Recruit diverse group

2. ***Investigate*** + Initial private estimate (Round 1)

3. Aggregate + Feedback

4. ***Discuss***

5. Revise ***Estimate*** (Round 2) + ***Aggregate***

Training materials

This document contains information about some of the important issues that impact how we understand outcomes in published papers. There are two main categories:

- statistical concepts commonly used in scientific papers; and
- information from scientific meta-research about the practice and publication of scientific findings, and previous replication studies

Both of these categories can be important for assessing research claims. For example, publication bias inflates the proportion of significant effects that are reported in the literature, and hence the number of Type-I errors.

Statistical concepts

As you read research claims for this project, you will encounter many different statistical tests and measures. We also refer to statistical concepts in our questions, e.g., "What is the probability that direct replications of this study would find a *statistically significant effect* [p<.05, two-tailed] in the same direction as the original claim?"

This section introduces some of the most relevant statistical concepts. It is far from exhaustive, and we identify some further resources if you'd like to know more.

The statistical material is mostly adapted from the MOOC provided by Daniel Lakens which can be found at: https://www.coursera.org/learn/statistical-inferences

P-values

p-values are widely used in the social and behavioural sciences, but they are also commonly misunderstood. Some common misconceptions include:

- The inverse probability fallacy falsely equating the probability of data, given the hypothesis (Pr (D|H) the conditional probability a p-value provides) with the probability of hypothesis, given the data (Pr (H|D) a different conditional probability)
- Replication fallacy the false belief that a p-value of .05 means that 95 times out of 100, the observed statistically significant difference will hold up in future investigations.
- Effect size fallacy falsely assuming a small p-value necessarily implies a large effect, when in reality, p-values are a function of both sample size and effect size, and so neither can be read directly from p.
- Clinical or practical significance fallacy falsely equating statistically significant with practical importance and/or clinical meaningfulness.

Things to remember:

The p-value does *not* tell you the probability that the null-hypothesis is true. Rather, it tells you about the probability of obtaining the data (or more extreme data) *given* the null hypothesis being true. It answers the question "how surprising are these data, or even more extreme data, under the assumption that there is no true effect?".

A p-value larger than 0.05 does *not* justify the conclusion that there is no (population) effect. It simply means the data is not surprising if the null hypothesis were true.

A couple of useful pointers:

- Generally speaking, you need larger samples to detect smaller effects.
- If there is no true effect (i.e. the null hypothesis is true) then every p-value is equally likely (a uniform distribution – refer to Daniel Lakens' video below). So, we have a chance of alpha of declaring a significant effect when in fact there is no effect.

If these concepts seem unfamiliar then please take a look at the following resources:

What is a p-value? (Video): This video by Daniel Lakens provides a refresher on how to interpret p-values, and warnings about how not to interpret them. 20 mins

Dance of the p-values (Video): This video by Geoff Cumming demonstrates the limitations of p-values in making inferences for replication. 11 mins

Type I and Type II Error

When we undertake a study, there are four types of observations we can make when we use significance tests:

- We find a significant effect where a true effect exists (true positive)
- We find a significant effect where no true effect exists (false positive or Type I error)
- We do not find a significant effect where no true effect exists (true negative)
- We do not find a significant effect where a true effect exists (false negative or Type II error).

You can learn more in this video by Daniel Lakens.

Type 1 and Type 2 errors (Video): Video by Daniel Lakens. 18 mins

Type 1 and Type 2 errors (spreadsheet to explore the effect of power, prior plausibility and alpha levels on Type 1 and Type 2 errors - refer to Daniel's video). You'll need to download this to play with it.

Effect sizes

Effect sizes can be a useful starting point for thinking about the practical significance of the results. (The effect might be statistically significant, but is it large enough to be meaningful?)

There are likely to be two main types of effect sizes you'll encounter in the research claims we're evaluating in the repliCATS project:

Unstandardized (or 'raw') effect sizes: measured on a scale that has a substantive meaning (e.g., number of milliseconds to respond, or the mean difference between two groups, measured in centimetres, etc.). An advantage of unstandardized effect sizes is that they can be understood on a substantive level. A disadvantage is that they are hard to evaluate without knowledge of the specific context, and hard to compare across studies that use

different scales.

Standardised effect sizes: There are a few types of standardized effect sizes. Measures like Cohen's d typically convey difference information in units of standard deviation (e.g., a mean difference divided by the standard deviation). Ostensibly this makes it easier to compare results across studies using different scales. But it can be harder to give meaning to them on a substantive level. And even though they are called 'standardized', it is important to remember that standard deviation is not a standard unit in the same way that kilograms or metres are.

Other units-free measures like correlations and variance explained convey strength of associations.

Effect sizes (Video): This video by Daniel Lakens provides a brief introduction to the importance of effect sizes. 10 mins

Cohen's d

A standardised effect measure that you are likely to encounter is Cohen's d. Cohen's d measures the standardized difference between the means of two groups, for example, the control group and a treatment group in an experiment. Formally, it is the difference between the group means (M_1 and M_2) divided by the standard deviation (SD, pooled from both groups or time points).

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

Cohen's d ranges from negative infinity to infinity. A Cohen's d close to zero means there does not seem to be much of an effect (but again, one should take context into account. Sometimes a small effect can still be very impactful). There can be much variation between Cohen's d values, depending on the design you use.

If you are unfamiliar with how to interpret effect sizes, there are useful benchmarks you can refer to. It is a subjective judgement, but benchmarks commonly given for Cohen's d are:

- 0.2 for small effects
- 0.5 for a medium effects
- 0.8 for a large effects

Please note that reliance on a benchmark to tell you whether the effect is 'small' or 'large' should only be a last resort- they are useful if you don't have anything better to go from. But the result will be more informative if you can relate it to comparable effects in the literature.

Cohen's d (Video): This video provides a brief introduction to Cohen's d. (8 mins).

Cohen's d (tool): The following tool can help you to explore Cohen's d

Correlation coefficients (r)

The correlation coefficient tells you about the relationship between two variables. Correlations range from -1 to +1, with 0 indicating no effect.

Again, benchmarks may be useful if you have no idea where to start, but be aware these benchmarks differ between domains and it's better to compare these to other similar studies:

- Small correlation is around 0.1,
- Medium correlation is around 0.3,
- Large correlation is around 0.5.

Correlations (Video): a useful introduction into correlations in the R family.

Correlations (tool): The following tool can help you to explore correlation coefficients

Eta squared: η^2, η^2_p and η^2_G

A common measure of effect size for variance in a model with multiple factors is η^2 (eta squared). Eta squared (η^2) describes the proportion of variance attributable to an effect standardised by the total variance across the sample. In particular, η^2 summarises the variance explained by one factor within an ANOVA design.

There are two variations of this measure for variance: η^2_p (partial eta squared) and η^2_G (generalised eta squared).

Partial eta squared (η^2_p) describes the proportion of variance that can be attributed to a particular factor after excluding variance explained by other factors in the model. In a one-way ANOVA, η^2_p and η^2 will be equal. However, in multiway or repeated measures ANOVA, will be larger than due to the variance explained by the additional factors.

Because η^2_p depends on which factors are actually measured in an experimental design, it will show different apparent effect sizes when some factors are measured in some designs but not measured in another. This makes it difficult to compare η^2_p across different studies. Generalised eta squared (η^2_G) was developed in order to avoid this issue, accounting for the variance associated with any measured, non-manipulated factors.

All of these effect size measures are useful in different scenarios, depending on the goals and design of the research.

- η^2 describes the variance attributable to a factor as a proportion of the total variance
- η^2_p describes the proportion of variance attributable to a particular factor after removing variance attributable to other factors in the model
- η^2_G describes the proportion of variance explained after excluding variance attributable to all manipulated variables but including all non-manipulated factors.

However, note that all of the three measures, η^2, η^2_p, and η^2_G are upwardly biased. That is, they tend to overestimate the population parameter, especially when the included sample size is small).

Epsilon squared ε^2 and Omega squared ω^2

Two other estimators for variance in a model with multiple factors that are less biased than η^2 are ε^2 (Epsilon squared) and ω^2 (omega squared). These estimators also have their partial and generalised equivalents. ε^2 and ω^2 are "essentially the same in practice", as they only

differ by an amount that will be negligible for most practical purposes. (Carroll & Nordholm, 1975, p. 544)

Both ε^2 and ω^2 are less biased than η^2, but have a higher level of variance (i.e., they do not tend to systematically under-estimate the population effect size but have a higher mean absolute error).

Statistical power

Power is a concept that is often poorly understood. It is formally defined as the probability of finding a statistically significant effect, assuming a certain study design and a given effect size. Hence, power depends on the effect for which it is calculated. This renders the power of a study a somewhat nonsensical concept: For the same study design, the power can be low to detect a tiny effect, but high to detect a large effect. Power is best understood as a property of the study design, and not of the data. Historically, the average statistical power of research published in psychology and other sciences has been low, which combined with publication bias, has led to replicability problems in the discipline.

Confidence intervals

A confidence interval (e.g. 95% CI) is an interval around a sample outcome. The 95% implies that if the study were repeated a large (infinite) number of times without bias, 95% of these intervals would include the population parameter.

What the likelihood is of another single sample mean (e.g. from a replication experiment) falling within this CI is a different question. This probability is called the Capture Percentage. It is a common misconception is to think of the Capture Percentage and the confidence percentage of the Confidence Interval as being the same, which would be true if and only if your original sample provided the exact population parameter. Of course, in practice, this is unlikely and unknowable. In general, the Capture Percentage is lower than the confidence percentage of the Confidence Interval. Consider the 95% CI from an initial study, and a replication study conducted without bias. On average, there's about an 83% chance that the initial CI will include the mean from the replication study.

There is a direct relationship between confidence intervals and tests of statistical significance. For example, if an effect is statistically different from 0 in a two-sided t-test with an alpha of 0.05, then the 95% CI for the mean difference between the two groups will never include 0; if 0 is in the CI, then the result is statistically non-significant. A typical short cut to reading statistical significance figures from CI error bars is to assume that if they don't overlap, then the results are statistically significant. This is an overly-conservative rule of thumb for alpha 0.05—in fact, 95% CIs can overlap by up to ~25% and still be statistically significant at 0.05.

We recommend the video provided below by Geoff Cumming for a simple explanation of Confidence Intervals.

Confidence Intervals (Video): Explains how to interpret confidence intervals.

Confidence intervals (Tool): This tool can be used to better understand confidence intervals.

Meta-research findings

Publication Bias

Publication bias may refer to the bias against publishing replication studies, that is, to journals having policies about accepting only 'original', 'new', 'groundbreaking' research.

Publication bias may also refer to the bias against publishing statistically non-significant, or negative, results. We usually use the term in this second way. This bias may come from editors and reviewers, i.e., be externally imposed. Or it may come from authors, self-selecting out of publishing non-significant results because of anticipated rejection. In reality, it is likely a combination of these factors.

Fanelli (2012) surveyed the percentage of positive (statistically significant) results in the published literature across a range of scientific disciplines. Most showed very high rates of statistically significant findings. Psychology--where over 90% of all published articles are statistically significant--was highest of all. This percentage is well beyond the expected rate of statistical significance, given the relatively low average statistical power found in other surveys of the psychology literature. Roughly speaking, the difference between the proportion of statistically significant results in the literature and the average statistical power of that literature, indicates the extent of publication bias. In heavily biased literatures, we should expect a higher than typical rate of false positives.

Registered Reports, and preregistration more generally, attempt to combat publication bias. Registered Reports differ from traditional papers in the sense that they are reviewed when only the study plan is known (that is, the introduction and the method section), and if their quality is deemed sufficient, they are published îndependent of the eventual outcome. So far Registered Reports do indeed seem to contain a great many more statistically non-significant results than typical publications (Scheel et al 2019, Allen and Mehler 2018).

Undisclosed flexibility (Questionable Research Practices)

In their excellent 2011 paper 'False Positive Psychology', Simons, Nelson and Simonsohn explain how practices like sampling extra data points, transforming variables, or excluding outliers *after* checking statistical significance and in the hope of meeting a particular threshold, increases the false positive rate and inflates effect sizes.

Questionable Research Practices (QRPs) range from extreme actions like fabricating data and fraud, to unwittingly employing transformations or tests in a way that inflates positive rates. When we talk about QRPs we mean to include practices like p-hacking, cherry picking, and HARKing. Whilst scientific fraud is indeed more common than we once thought, it is still quite rare. These other QRPs, however, are surprisingly common, as repeated self-report surveys have shown (John et al 2012, Agnoli et al 2017, Fraser et al 2018).

Replication rates

Over the last 4-5 years, large scale replication studies have on the whole found low replication rates (despite having designs with a high power for finding the original, or smaller, effect size). The first of these--The Reproducibility Project Psychology (RPP)--obtained a statistically significant result in the same direction as the original study in only 36% of cases

(i.e., in 35/97 replications). Moreover, the effect size was on average about half the size of the original studies (OSC, 2015)

Since then there have been several others studies of this kind, in economics (Camerer et al 2016), personality psychology (Soto et al 2019), experimental philosophy (Cova et al 2018), cancer biology (Errington et al 2014), and in social science papers published in Science and Nature (Camerer et al, 2018). There have also been multiple rounds of Many Labs projects, where as the name suggests, several independent researchers repeat the same study (e.g., Many Labs 1 Klein et al, 2014). Combined, these replication studies suggests that the current body of literature is, on average, overestimating the proportion and size of population effects.

Plausibility

To assess a certain study, it is important to take the plausibility of the effect being true into account as well. It makes sense to take prior knowledge or prior expectations into account: seldom does one start completely blank when reading a study. Thus, someone may be confident of a certain effect to replicate for study A, but the same person could be justifiably unconvinced for the exact same results for another study. Basically, this could be summarized in five words: Extraordinary claims require extraordinary evidence.

Note that this point has also been highlighted in the list of 'things to consider' in the Guide to Round 1. Background ideas about plausibility, where they have been informed by prior experience and evidence, can be a very important factor.

As we said at the beginning of this document, this isn't an exhaustive list of things that might help inform your estimates, but hopefully you will find the other links and references useful.

If you feel there is something crucial we have missed / overlooked please contact us repliCATS-contact@unimelb.edu.au

Want to know more?

The above sections aim to provide a few simple tools regarding some common statistics you might encounter and inferences you might be inclined to make. Much of the statistical material is modelled from Daniel Laken's Coursera course, which we recommend if you feel that you need a complete refresher.

We could not cover all of the statistics you are likely to encounter. However, below we have included some useful videos that may help you in predicting replication.

Additional tutorials:

Frequentism, Likelihoods, Bayesian Statistics (Daniel Lakens- 9min)

Confidence intervals (Daniel Lakens- 12min)

Sample size justification (Daniel Lakens – 11min)

P-curve analysis (Daniel Lakens 9 min)

Replications (Daniel Lakens- 13min)

Likelihoods (Daniel Lakens- 16min)

Binomial Bayesian Inference (Daniel Lakens – 14min)

Bayesian Thinking (Daniel Lakens – 11min)

Useful tools:

Some useful visualisation tools: https://rpsychologist.com/tag/visualization

APA statistical notation (Sussex University guide). here

References

Agnoli F., Wicherts J. M., Veldkamp C. L. S., Albiero P. and Cubelli R. (2017) Questionable research practices among italian research psychologists. PLOS ONE 12, e0172792, https://doi.org/10.1371/journal.pone.0172792

Allen, C. and Mehler, D. (2018): Open Science challenges, benefits and tips in early career and beyond. Preprint: https://psyarxiv.com/3czyt/

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016). Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433–1436 ; https://science.sciencemag.org/content/351/6280/1433 ; https://osf.io/bzm54/

Camerer C. F., Dreber A., Holzmeister F., Ho T.-H., Huber J., Johannesson M., Kirchler M., Nave G., Nosek B. A., Pfeiffer T., Altmejd A., Buttrick N., Chan T., Chen Y., Forsell E., Gampa A., Heikensten E., Hummer L., Imai T., Isaksson S., Manfredi D., Rose J., Wagenmakers E.-J. and Wu H. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2, 637-644. https://www.nature.com/articles/s41562-018-0399-z

Carroll, R. M., & Nordholm, L. A. (1975). Sampling Characteristics of Kelley's ε and Hays' ω. Educational and psychological measurement, 35, 541-554. doi:10.1177/001316447503500304

Chang, Andrew C., and Phillip Li (2015). “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ”Usually Not”,” Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.083

Cova F., Strickland B., Abatista A., Allard A., Andow J., Attie M., Beebe J., Berniūnas R., Boudesseul J., Colombo M., Cushman F., Diaz R., N’Djaye Nikolai van Dongen N., Dranseika V., Earp B. D., Torres A. G., Hannikainen I., Hernández-Conde J. V., Hu W., Jaquet F., Khalifa K., Kim H., Kneer M., Knobe J., Kurthy M., Lantian A., Liao S.-y., Machery E., Moerenhout T., Mott C., Phelan M., Phillips J., Rambharose N., Reuter K., Romero F., Sousa P., Sprenger J., Thalabard E., Tobia K., Viciana H., Wilkenfeld D. and Zhou X. (2018)

Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*.

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. Perspect. Psychol. Sci. 3, 286-300 (2008). doi:10.1111/j.1745-6924.2008.00079.xpmid:26158948

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? Psychological Methods, 11(3), 217-227.

Errington T. M., Iorns E., Gunn W., Tan F. E., Lomax J. and Nosek B. A. (2014) An open investigation of the reproducibility of cancer biology research. *eLife* **3**.

Fanelli, D. (2012) Negative results are disappearing most disciplines and countries Scientometrics 90, 891–904, DOI:10.1007/s11192-011-0494-7

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F (2018) Questionable research practices in ecology and evolution. PLoS ONE 13(7): e0200303. https://doi.org/10.1371/journal.pone.0200303

Herrera-Bennett, A. (2019). How do researchers evaluate statistical evidence when drawing inferences from data? (Unpublished doctoral dissertation). Ludwig-Maximilians-Universitaet, Munich, Germany. https://osf.io/ndum8/

John L. K., Loewenstein G. and Prelec D. (2012) Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. Psychological Science 23, 524-532. DOI: 10.1177/0956797611430953

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B.A. (2014). Investigating variation in replicability: A “many labs” replication project. Social Psychology, 45(3), 142- 152. http://dx.doi.org/10.1027/1864-9335/a000178

Maxwell, S.E. (2004) The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. Psychological Methods 9, 147-163

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science 349 (6251), 10.1126/science.aac4716. https://osf.io/ezcuj/wiki/home/

Scheel, A., Schijen, M. and Lakens, D. (2019) Positive result rates in psychology: Registered Reports compared to the conventional literature, Eindhoven University of Technology https://osf.io/854zr/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singleton Thorn, F. (in prep). PhD thesis, University of Melbourne. https://osf.io/h8u9w/

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. Psychological Science, 30(5), 711–727. https://doi.org/10.1177/0956797619831612

Yong, E. (2012) Replication studies: Bad copy. Nature (2012). https://www.nature.com/news/replicationstudies-bad-copy-1.10634