Heterogeneity in close and conceptual replications

AUDREY LINDEN, JOHANNES HÖNEKOPP
OPEN SCIENCE, MARCH 2019
Context

- Replication crisis

- Replication projects
  - Close replications (Many Labs, Registered Replication Reports)
  - Keep design, materials, analyses, etc. as close to original as possible

- Meta-analyses
  - Conceptual replications
  - Different designs, materials, participants, analysis, etc.

- Heterogeneity – how much true effect sizes differ across studies
Why care about heterogeneity?

- **Affects statistical power** (McShane & Böckenholt, 2014; Shrout & Rodgers, 2018)
  - Heterogeneity decreases power
  - Power calculations should take this into account. But first we need a reliable estimate of heterogeneity, and what may be driving this.
  - Could this explain low success in 100 close replications from Open Science Collaboration (2015)?

- **Design of practical applications**
  - Heterogeneity can tell us something about the level of certainty around the result of the ‘next study’
  - Successful translation of research into practice depends on consistency of findings
Aims

- Derive an estimate of heterogeneity in close replication studies
- Compare this to heterogeneity in a large sample of meta-analyses
- Investigate some possible causes of heterogeneity
Hypotheses

- Heterogeneity in close replications expected to be low

- Heterogeneity in conceptual replications
  - Higher in social than cognitive psychology
    - Higher replication success for cognitive than social psychology (Open Science Collaboration, 2015)
  - Higher in social than organisational psychology
    - Higher correlation between lab and field studies in organisational than social psychology (Mitchell, et al., 2012)
Methods

- 40 close replication studies (Many Labs and Registered Replication Reports)
- 147 meta-analyses sampled (cognitive, organisational, social psychology)
  - Cohen’s $d$ as measure of ES
  - $\tau$ as measure that quantifies heterogeneity
    - Generally assumed that population ES for a given phenomenon follow a normal distribution
    - $\tau$ is their standard deviation
- $d$ and $\tau$ calculated by re-doing all meta-analyses
Methods – Two approaches to moderators

- 25 meta-analyses with \( k \geq 60 \), and with sufficient information to re-examine moderator analyses
  - Excluded ‘broad’ subsets (e.g. adults, children, mixed – mixed sample excluded)

- For meta-analysis as a whole, rated broadness/narrowness of inclusion criteria for studies on a 5-point scale
  - E.g. is the question addressed narrow/broad
  - Does manipulation of IV/DV follow standard protocol
How do meta-analyses address heterogeneity?

Out of 147 meta-analyses...

- 54% reported a measure of heterogeneity
- Heterogeneity quantified in only 38 cases (26%)

Post-PRISMA? (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, 2009)

- Heterogeneity only reported in 60% of cases
Findings in our sample

- Average $\tau$ was low ($M = 0.08$) in close replications.

- Average $\tau$ was much higher ($M = 0.33$) in conceptual replications.

- Overall ES for average close replication was $d = 0.24$; for meta-analysis this was $d = 0.45$.

- Heterogeneity in conceptual replications:
  - No significant differences between the 3 sub-disciplines (cognitive, social, organisational).
  - The distinctive success rates of these sub-disciplines in terms of replication is not reflected in heterogeneity levels.
What does this level of heterogeneity mean?

- Average $\bar{T} = 0.33$
- Cohen's $d$ of 0.2/0.5/0.8 are often used as benchmarks for small/medium/large effects
- All of these occur frequently in the distribution of true effect sizes
- Can expect replications to find results in the opposite direction
Close replications

- Average $\bar{T} = 0.08$
- High consistency in results
Variability in heterogeneity – why?

- Mixing apples and oranges: broad versus narrow inclusion criteria

- Looked at moderators in a sub-set of 25 large meta-analyses
  - Looked at one moderator in each case

- No significant difference in heterogeneity between overall meta-analyses ($M = 0.34$) and subset based on moderators ($M = 0.36$)

- Broad/narrow inclusion criteria?
  - Narrow sub-sample, heterogeneity still high ($M = 0.29$)
Variability in heterogeneity – why?

Explanatory analyses

- Research areas with larger ES have greater heterogeneity (Kenny & Judd)
  - Strong relationship between mean $d$ and $\tau$
  - For close replications ($r = .70, p < .001$)
  - For conceptual replications ($r = .45, p < .001$)

- Maturity of a research field, or broader inclusion criteria?
  - Relationship between $k$ and $\tau$ ($r = 0.30, p < .001$)

- Establishing an effect -> exploring boundaries
  - Used a median date split
  - No significant difference in $\tau$ between the earlier and later dates
Conclusions

- Heterogeneity in close replications proved low – producing reliable results is possible.

- Heterogeneity in close replications reduces power only marginally - for sample sizes that generate 80% power at zero heterogeneity...
  - For $\tau = 0.08$, power
    - Does not reduce for large effects
    - Drops to 78% for medium effects
    - Drops to 71% for small effects

- For Open Science Collaboration (2015), mean effect size was large ($d = 0.87$)
  - Power therefore not affected
  - ‘Hidden moderators’ typically of no concern
Conclusions

- Heterogeneity in meta-analyses is large (and not strongly affected by bias)
  - Mean ES reported in MA with large heterogeneity have limited use

- Research Planning
  - Difficult to estimate efficacy of an intervention (effect could be in opposite direction)
  - Heterogeneity and effect size determine how predictable the result of the ‘next study’ is
Conclusions

- $\varphi = 0.33$ has a more dramatic effect on power
  - Drops to 71% for large effects
  - Drops to 66% for medium effects
  - Drops to 57% for small effects
Implications

- Cumulative knowledge
  - Science = quest to explain apparent complexity in observations through simpler fundamental principles
  - (Unexplained) heterogeneity is a measure of how much this quest fails

- Falsifiability of theories
  - Say test of theory $X$ requires induction of good mood. We use mood induction procedure $Y$
  - When effectiveness of $Y$ is debatable (large heterogeneity), failed test of theory $X$ becomes meaningless
  - Weak tools undermine falsification and thereby good theoretic progress

- When knowledge $Y$ is used as a tool, we need to replicate as closely as possible
Limitations

- Difference in effect size between close replications ($d = 0.24$) and conceptual replications ($d = 0.45$)
  - How does low heterogeneity in close replications generalise to psychological research findings?

- We used Hunter-Schmidt meta-analysis model
  - Similar results for Hedges, and DerSimonian-Laird models
  - HS estimates of heterogeneity were slightly more conservative
Thank you

- Questions...