1st Paul Meehl Graduate School PhD Day

20th September 2024, TU/e Eindhoven

Conference website: https://paulmeehlschool.github.io/workshops/second%20year/phdday/

Program

Time	Activity
9:00 - 9:15	Opening and Welcome
9:15 - 9: 45	**Jack Fitzgerald** The Need for Equivalence Testing in Economics
9:45 - 10:15	**Cas Goos** Mapping Reproducibility Barriers, Root Causes, and Interventions in the Psychology Meta-Research Literature, a Systematic Review
10:15 - 10:45	**Iris Willigers** Scaling problems in Raven's Progressive Matrices after Rasch Analysis: A Simulation Study and Multilevel Meta-Analysis
10:45 - 11:15	Coffee Break
11:15 – 11:45	**Raphael Merz** The Prevalence of Nonsignificance Misinterpretations in Psychology, and its change over time

11:45 – 12:15	**Tim Mori** Is uncertainty in clustering algorithms sufficiently addressed in medical research? A practical example from diabetes research
12:15 -12:45	**Martin Buchner** Settling Settler Mortality: An Expert Survey on the Replication Debate Between Acemoglu et al. (2001) and Albouy (2012)
12:45 - 13:45	Lunch Break
13:45 - 14:45	**Plenary discussion** **Career in Metascience** Leo Tiokhin - Miguel Silan - Olmo van den Akker - Tamarinde Haven
14:45 – 15:15	**Talk about PYMS** Olmo van den Akker
15:15 – 15:45	Coffee Break
15:45 - 16:15	**Finn Luebber** Improve Theorizing and Value of Empirical Work with an Interactive Tool for Data Simulation
16:15 - 16:45	**Dwayne Lieck** Contextualizing Effects in Educational Research
16:45 - 17:15	**Cristian Mesquida** Validity and Reproducibility of Pre-study Power Analysis
17:15 - 17:45	Closing talk
17:45 – 21:00	Dinner

Abstracts (in alphabetical order)

Cas Goos

Mapping Reproducibility Barriers, Root Causes, and Interventions in the Psychology Meta-Research Literature, a Systematic Review

Widespread concern over the fact that the results in many scientific articles cannot be reproduced using the same data and analysis strategy as used originally has resulted in numerous initiatives to investigate and improve the reproducibility of scientific articles. In line with these initiatives, we plan to systematically review the meta-research literature on reproducibility in psychology in the last 15 years to map the most important issues and interventions for reproducibility identified in this literature. To map these elements together, we created a schematic workflow of the ideal scenario when reproducing an article that we will use to annotate any barriers that might prevent an article being reproduced. We will then annotate root causes to these barriers, as well as possible solutions to remove the barriers and reduce the impact of the root causes. We plan to synthesize the annotated pieces of texts from single articles into overarching elements representing common barriers, root causes, and solutions as described in the literature using qualitative methods while supplementing this with quantitative information on the estimated impact of the elements on each other, and whether journals can implement the suggested solutions. Together, this information will be used to identify solutions that journals can implement to effectively address important barriers and structural root causes diminishing the reproducibility of psychological research. We are currently still refining the article retrieval, screening, and coding protocols and believe this conference would be an excellent opportunity to get feedback from peers.

Cristian Mesquida

Validity and Reproducibility of Pre-study Power Analysis

Researchers should design and conduct studies that have a high chance of finding their effect of interest assuming there is one to be found. One way to achieve this is by conducting a pre-study power analysis to determine the sample size required to achieve a desired level of statistical power given an effect size, statistical test and alpha level. Despite the importance of statistical power at the design stage, published studies in sport and exercise science often employ sample sizes that might be too small to detect the effect size of interest. Studies with underpowered designs can increase the rate of false positives and false negatives, giving rise to potentially misleading scientific literature. Therefore, we sampled 350 articles published across 10 journals in the field of sport and exercise science and used a coding form to (1) estimate how many published studies report using a pre-study power analysis; (2) assess their reporting practices; (3) analyze if the results are reproducible; and (4) assess how often these analyses use

G*Power's default option for mixed-design ANOVAs—which can be misleading and yield sample sizes that are too small for a researcher's intended purpose.

Dwayne Lieck

Contextualizing Effects in Educational Research

Most researchers and practitioners will argue that knowing which methods and technologies for instruction best foster learning is of high practical relevance. Still, the methods researchers in education use to interpret the magnitudes of their effects are basic – often simply citing conventions like Cohen (1988) or Hattie (2009). Whilst many researchers would agree in seeing these solutions as inadequate, little time is available to do more. Especially researchers going into classrooms to do research have to put enormous effort into conducting their studies, making effect magnitude interpretation a lower priority. This is especially problematic, as many "medium" or "large" effects deflate when transferred from the lab to the classroom, leading to highly underpowered studies. While an easy-to-use, more advanced general framework to interpret effects' magnitudes is an ideal to strive towards, any improvement on this topic is important to advancing educational research. I therefore start with a discussion and critique of current practices for interpreting the magnitude of effects, which leads into the construction of a framework for effect contextualization in educational research. This framework contains four categories under which effects can be contextualized: 1) Didactic Setting, 2) Material Types, 3) Time Frame and 4) Comparison. While this contextualization is insufficient to fully judge the practical relevance of effects, it gives more information for researchers and practitioners alike to evaluate effects with minimal time invested.

Finn Luebber

Improve Theorizing and Value of Empirical Work with an Interactive Tool for Data Simulation

Data simulations are an important and still underused tool in empirical research, facilitating power analyses, checking the robustness of statistical models and the effect of unreliable measures, or evaluating the influence of potential unmeasured variables. However, existing software packages require quite advanced programming skills and are thus not accessible to all researchers. Thus, we developed a ShinyApp which enables users to construct models as a directed acyclic graph (DAG) in an intuitive interface from which data can be directly simulated, analyzed and the results compared to the theoretical model.

Within the model, users can add and adjust variables and causal connections between them, as well as setting error magnitudes and measurement precision. Based on existing R packages, the app outputs adjustment sets (variables to adjust for to identify the effect of interest) and simulates data from the model, which is then analyzed within the app. Comparison of the

simulated result distribution with the theoretically postulated causal effect size can directly reveal expected bias in and precision of the statistical effect estimate (including power) in the researcher's suggested analysis strategy.

Thus, firstly, this tool will direct users to formalize verbal theories by forcing quantitative parameters for variables and their causal connections. Secondly, the implications of this formalization in terms of precision, power, and bias are directly fed back, transmitting important information prior to actual empirical work, also facilitating and predicting replicability efforts in the long run.

Iris Williger

Scaling problems in Raven's Progressive Matrices after Rasch Analysis: A Simulation Study and Multilevel Meta-Analysis

The Raven's Standard Progressive Matrices exhibits Flynn Effects over time, potentially leading to ceiling effects and unreliable test outcomes. Our simulation studies explored how increasing mean levels of true (fluid) intelligence affect various outcomes, including mean sum scores, (corrected) standard deviations, true reliability, KR-20, KR-21 and (corrected) observed Cohen's ds. We assessed the effect of different conditions of latent mean (differences) to assess item difficulties under the Rasch model with input of item responses of an existing dataset. Rasch Analysis revealed no major issues in measuring (fluid) intelligence. However, the sum score distributions did suffer from floor and ceiling effects. True reliability and reliability estimates were low for low and high latent means. In addition, observed Cohen's d was biased compared to latent Cohen's d, and corrections with the KR-20 or KR-21 did not decrease bias for all conditions. In addition, shorter test versions of the test yielded even lower reliability estimates. We also examined meta-analytical implications for intelligence interventions on between-groups Cohen's d measured by the Raven's Standard Progressive Matrices. There was no significant mean meta-analytic effect found, with also no moderating effects of KR-21 and number of items. There was a positive relationship between the number of items and KR-21
for these studies. In conclusion, our study shows that the measurement precision of the Raven's Standard Progressive Matrices is inadequate for making inferences on the sum score distribution and Cohen's d of samples in the lower or higher latent means.

Jack Fitzgerald

The Need for Equivalence Testing in Economics

I introduce equivalence testing procedures that can provide statistically significant evidence that economic relationships are practically equal to zero. I then demonstrate their necessity by systematically reproducing the estimates that defend 135 null claims made in 81 articles from top economics journals. 36-63% of these estimates fail lenient equivalence tests. Though prediction platform data reveals that researchers find these equivalence testing failure rates (ETFRs) to be unacceptably high, researchers actually anticipate unacceptably high ETFRs, accurately predicting that ETFRs exceed acceptable thresholds by around 23 percentage points. To obtain ETFRs that researchers deem acceptable, one must contend that nearly 75% of published effect sizes in economics are practically equal to zero. This implies that Type II error rates are unacceptably high throughout economics. This paper provides economists with empirical justification, guidelines, and commands in Stata and R for conducting credible equivalence testing and practical significance testing in future research.

Martin Buchner

Settling Settler Mortality: An Expert Survey on the Replication Debate Between Acemoglu et al. (2001) and Albouy (2012)

This study investigates whether experts reach consensus over a famous replication debate. It focuses on the highly influential paper by Acemoglu, Johnson, and Robinson (2001) that examines the impact of colonial institutions on long-term economic development. A critical comment by Albouy (2012) challenged the original findings, prompting a rebuttal from Acemoglu et al. As both sides stick firmly to their positions, the debate remains unresolved.

To address this lack of consensus, we conduct an expert survey targeting academics across various fields. Participants are recruited through multiple strategies, including identifying scholars who cited at least one of the debate papers, published papers critical of the empirical approach used by the original study, or employed a similar methodology. Using a structured online questionnaire, we gather 352 fully completed responses, primarily from economists, with the majority holding positions as professors or associate professors.

We find that experts slightly lean towards the replicator's side, though there is no overall consensus, as indicated by the scattered distribution of responses. In addition to this primary finding, we observe that initially, experts are more familiar with the original paper than with the comment. However, reading summaries of all three debate papers eventually led them to change their priors, making them less convinced by the original study. Seniority and professional expertise are associated with increased support for Albouy's critique, suggesting that more experienced academics tend to be more critical of the original study's conclusions.

Raphael Merz

The Prevalence of Nonsignificance Misinterpretations in Psychology, and its change over time

Numerous studies confirm that researchers frequently misinterpret key statistics in published psychology articles. A particularly prevalent issue identified by previous research is the tendency of researchers to misinterpret nonsignificance as representing no true effect (estimated at over 60% of published psychology articles reporting a nonsignificant finding). Nevertheless, methodological decisions mean this meta-science research likely failed to accurately capture the real prevalence rate. Also, related meta-science efforts have yet to examine whether researchers are less likely to make this interpretative error today than they were many years ago (when researchers were less educated on the issue). Accordingly, the present study aims to investigate these points – to clarify the prevalence of nonsignificance misinterpretations in published psychology articles, to examine whether this issue has improved over the past decade, and to explore whether researchers generally know that nonsignificance does not reflect an effect's absence. To achieve these aims, we looked at nonsignificance statements in the discussion sections of 599 articles across three time points (2009, 2015, 2021) from ten psychology journals of varying impact factors. We then coded each statement as correctly or incorrectly interpreting nonsignificance, and whether incorrect interpretations were sample-based (e.g., ‘age did not affect our self-control’) or population-based (e.g., ‘age does not affect self-control’). Preliminary results reveal a higher prevalence of these misinterpretations compared to prior studies (80% incorrect: 60% population-based and 20% sample-based), with only minor (descriptive) changes over time points and differences across journals, further highlighting the need for better education.

Tim Mori

Is uncertainty in clustering algorithms sufficiently addressed in medical research? A practical example from diabetes research

Background and aims: In recent years, there has been a lot of hype in diabetes research about the so-called “novel diabetes subtypes”, which are data-driven clusters that were derived using standard clustering algorithms. Assigning people with diabetes into discrete clusters based on their clinical features is appealing for clinicians. However, the robustness and reliability of such cluster assignments has received very little attention in the diabetes research community. This study aimed to highlight the challenges of discrete cluster assignments and the importance of considering how certain we are in a person’s cluster assignment.

Methods: We developed a simple graphical tool for doctors and researchers to visualize the certainty in the cluster assignment of individuals with diabetes. Moreover, we propose an easy-to-interpret statistical measure to quantify this uncertainty. Using data from the German

Diabetes Study (GDS), we compare the cluster uncertainty across the different diabetes clusters.

Results: The graphical tool revealed that some individuals with diabetes differed substantially from the typical profile of their assigned cluster, limiting the utility of the clusters for these individuals. Our proposed uncertainty measure showed that, on average, there is more certainty in the assignment to some diabetes clusters compared to others.

Conclusion: Though this is currently not done in practice, clustering of people with diabetes should always be accompanied by a measure of cluster uncertainty. More broadly, drawing attention to the limitations of current analysis approaches and providing practical tools can enhance research rigor and improve the reliability of findings in medical research.