6th Perspectives on Scientific Error Workshop

February 29th–March 2nd, 2024, TU/e Eindhoven

Conference website:
https://www.eurandom.tue.nl/perspectives-on-scientific-error-6th-edition-29-february-1-march-2024-tu-e-eindhoven/

Program

Day 1: Thursday, February 29th

Location: MetaForum building room 11 and 12

8:30 - 9.10	*Registration*
9:10 - 9:15	*Welcome and Opening*
9:15 - 9:45	**Olmo van den Akker, Marcel van Assen, Marjan Bakker, & Jelte Wicherts** Preregistration: Past, Present, and Prospects [slides]
9:45 - 10:15	**David F. Urschler** Do we really know what we are talking about? The prevalence of well- and ill-defined psychological constructs in 150 meta-analytic reviews. [video]
10:15 - 10:45	**Ana Martinovici** Is this real data? How to detect data fabrication in Qualtrics questionnaires [video]
10:45 - 11:15	*Coffee break*
11:15 - 12:30	**Keynote 1: Catarina Dutilh Novaes** Mistakes in Mathematical Proofs [video] [slides]
12:30 - 14:00	*Lunch break*

14:00 - 14:45	**Rink Hoekstra, Nina Schwarzbach, Henk Kiers, James Steele, & Fiona Fidler** The Illusion of Objectivity - Discussing the Hidden Subjectivity in Quantitative Social Science [video]
14:45 - 16:00	**Keynote 2: Ian Hussey:** The ERЯOR project: A three-prong effort to improve post publication critique and error detection [video] [slides]
16:00 - 16:45	*Coffee break*
16:45 - 17:30	**Fiona Fidler** Can forecasts of replicability improve peer review? [video] [slides]
19:00	*Dinner* *(Walk-in for drinks from 18:00)*

Day 2: Friday, March 1st

Location: MetaForum building room 11 and 12

9:15 - 9:45	**Candida Sánchez Burmester** Travelling claims: changing presentations of a nanoparticle in different genres
9:45 - 10:15	**Veronika Cheplygina, Amelia Jiménez-Sánchez, Gaël Varoquaux** Shortcuts and other shortcomings in machine learning for medical imaging [video] [slides]
10:15 - 10:45	**Gerit Pfuhl, Adrian Helgå Vestøl, Ole Fredrik Borgundvåg Berg** Do Norwegian publishing incentives lead to questionable research practices? [video]
10:45 - 11:15	*Coffee break*
11:15 - 12:30	**Keynote 3: Judith ter Schure** Making research Evidence-Based ruins error control from the sampling space
12:30 - 13:30	*Lunch break*
13:30 - 14:15	**Aurélien Allard**

	Theory building and replicability: on the value of basic facts without theoretical foundations [video] [slides]
14:15 - 15:00	**Christian Hennig** Understanding statistical inference based on models that aren't true [video] [slides]
15:00 - 15:30	*Coffee break*
15:30 - 16:45	**Keynote 4: Simine Vazire** The dog that caught the car: Turning scientific values and principles into journal policy and practice
16:45 - 18:00	**Poster Session** • **Andrea Kis, Elena Mas Tur, Daniël Lakens, Krist Vaesen, Wybo Houkes** Leaving academia: PhD attrition and unhealthy research environments • **Auste Valinciute** Correction of scientific errors in the media • **Bente Sinke, Matteo Colombo, Michal Klincewicz** Scientific credit and the Matthew effect in neuroscience • **Cristian Mesquida** What is your research hypothesis? On the importance of deconstructing your research hypothesis to improve the severity of hypothesis tests • **Jamie Cummins** Erring on the side of caution: Family-wise Type II error • **Ligaya Breemer, A. E. van 't Veer, L., P. M. Isager, T. Heyman, T. van Leeuwen, & M. C. Makel** The Prevalence of Replications in Psychology Revisited • **Peter Stilwell, Tom Heyman** Sequential Strategies: Assessing Group Sequential Designs with Real Data • **Raphael Merz, Stephen Lee Murphy, Linda-Elisabeth Reimann, Aurelio Fernandez Zapico** The Prevalence of Nonsignificance Misinterpretations in Psychology, and its change over time

	• **Taym Alsalti** Meta-analysis: Is There Still Hope? • **Leonhard volz** PsychoModels: Giving strong theory building a FAIR chance
19:00	*Dinner*

Day 3: Saturday, March 2nd

Hackathon: Reducing Scientific Error in Practice

Location: MetaForum building room 11 and 12

10:00 - 10.15	*Arrive, coffee*
10.15 - 10.25	**Welcome** Simine Vazire & Tom Hardwicke
10:25 - 11:25	Hackathon pitches (10 minutes each) 1. **Risk of Statistical Error (ROSE)** Rickard Carlsson & Natalie Hyltse 2. **Automating checks for reporting standards, statistical inferences, and open science practices** Daniël Lakens 3. **Assessing Computational Reproducibility** Lisa DeBruine 4. **Self-correcting science: Increasing the discoverability and usability of post publication scrutiny and error detection tools** Ian Hussey 5. **The garage is open – where are the cars? How do we coordinate quality control work in practice?** Peder Isager & Anna van 't Veer
11.25 - 12.25	**Start Hackathons**
12.25 - 13.25	*Lunch break*
13.25 - 17.00	**Hackathons**
17.00 - 17.30	**Wrap-up: share hackathon results, plans going forward.**

See the Hackathon abstracts at the end of the document.

Keynote Speakers

Catarina Dutilh-Novaes

VU Amsterdam

I am a professor and University Research Chair at the Department of Philosophy of the VU Amsterdam. I am also a Professorial Fellow at Arché in St. Andrews (2019-2024). I am currently running the ERC Consolidator project 'The Social Epistemology of Argumentation' (2018-2023). I am an associate editor for Analysis, and a member of the Royal Netherlands Academy of Arts and Sciences (KNAW).

My monograph The Dialogical Roots of Deduction won the 2022 Lakatos Award.

My main fields of research are history and philosophy of logic, philosophy of mathematics, and social epistemology. I also have general interests in medieval philosophy, philosophy of psychology and cognitive science, general philosophy of science, philosophy of mind, philosophy of technology, issues pertaining to gender and race, and empirically-informed approaches to philosophy in general.

Simine Vazire

University of Melbourne

My research examines whether and how science self-corrects, focusing on psychology. I study the research methods and practices used in psychology, as well as structural systems in science, such as peer review. I also examine whether people know themselves, and where our blind spots are in our self-knowledge. I teach research methods.

I am editor in chief of Psychological Science (as of 1 Jan, 2024) and co-founder (with Brian Nosek) of the Society for the Improvement of Psychological Science.

Ian Hussey

University of Bern

My research examines the robustness of research claims in psychology, particularly those related to poor measurement, measurement flexibility, and computational reproducibility. I study the contingencies that govern scientists' behavior and our processes of scientific knowledge production.

Currently, I am particularly interested in post publication critique and error detection. Together with Malte Elson and Ruben Arslan, I run the ERЯOR (Estimating the Reliability and Reproducibility of Research) project: A bug bounty program for science.

Judith ter Schure

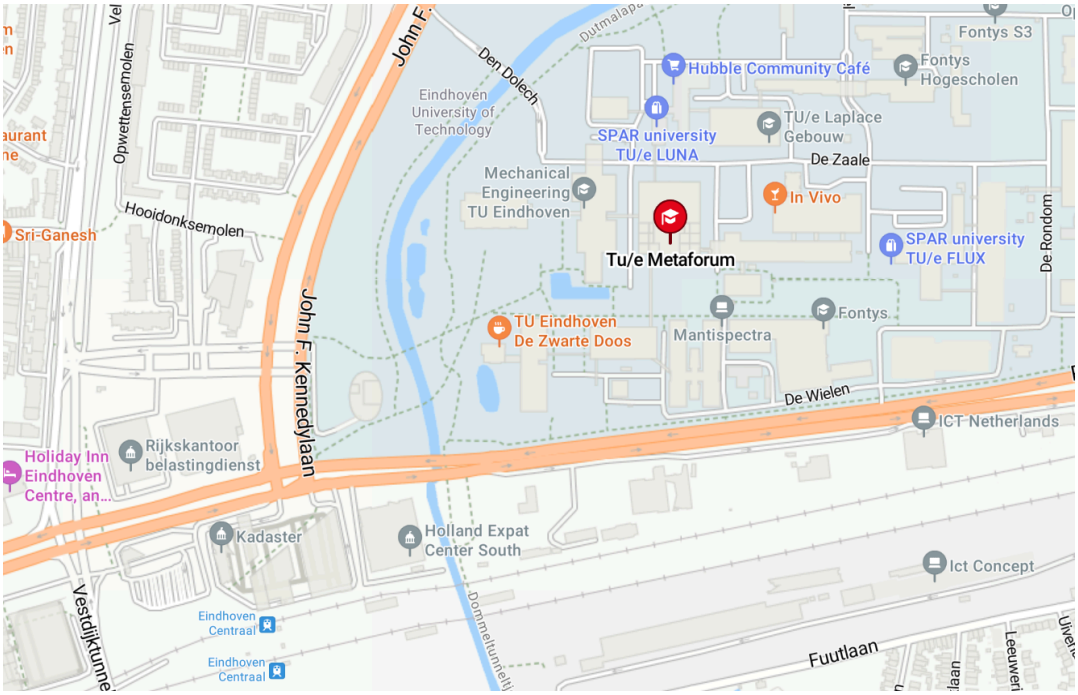
Amsterdam UMC

Judith ter Schure's interests lie in foundations of statistics as well as in applied work. She divides her time between research and consultancy, both on randomized clinical trials (RCTs) (before at CWI and significanthelp.nl, now at Amsterdam UMC). Her general motivation is the effect of statistics on society, which also inspires occasional writing for a wide audience – previously published by De Correspondent and the Dutch Journal of Medicine – and participation in popular science events like Nacht van de Wetenschap. Judith is a former member of the daily board (treasurer, 2019-2023) of the Netherlands Society for Statistics and Operations Research (vvsor.nl).

Her research is on ALL-IN meta-analysis: new approaches to meta-analysis that are Anytime, Live, Leading and possibly on INterim data from RCTs. ALL-IN meta-analysis can be applied retrospectively as well as prospectively, to evaluate the evidence once or sequentially. Because the intention of the analysis does not change the validity of the results, the results of the analysis can change the intentions. Any ALL-IN meta-analysis can be turned into a living one, or even become 'live' or 'real-time' by including interim data from trials that are still ongoing.

Practical Information

Location: Eindhoven University of Technology, MetaForum building, rooms MF 11 and MF 12.



Conference Dinners

Februrary 29th. O'Shea's Irish Pub. (~7:00pm)

Indian Buffet, 18 euro
Jan van Lieshoutstraat 9, 5611 EE Eindhoven

March 1st. Hubble Community Café. (~7:00pm)

Vegetarian Buffet, 9.50 euro
On campus: De Lampendriessen 31-05, 5612AH Eindhoven

- **Menu**: We notified the restaurants about your dietary restrictions and preferences, and they will have options for you.
- **Drinks**: The payment covers the food. You can order/pay for drinks at the venue.
- **Payment**: Daniel will be collecting the payments. You can use PayPal or pay by cash. Locals can pay via Tikkie. We would appreciate it if you paid before the dinners (link).
- **Proof of attendance**: We will provide you with a document acknowledging your participation and payment in the dinner, should you require it for reimbursement.

Accommodation

We can recommend the Social Hub Eindhoven. If you book there before January 18th, you can take use of their winterdeals and get 25% off. Please note these bookings are non refundable. If you book later, you can get a 15% discount with the promo code **LEIDEN-TSH** on their website. This promo code gives **15%** discount on the best available (flexible) hotel rates and is available for a stay between **27-02-24** and **04-03-24**.

A more affordable option is: Boutiquehotel Sycamore

Travel Information

You can use a credit card or a phone set up for payment to tap in and out for all public transport across the Netherlands. Alternatively, you can buy train tickets via www.ns.nl.

Door-to-door journey planner (www.9292.nl).

Journey planner NS (Dutch Railways) (www.ns.nl)

Eindhoven University of Technology is a 5-minute walk from Eindhoven Central Station.

Code of Conduct

At and during the conference, we follow the Code of Conduct of the International Network of Open Science Communities (find it here). We expect all participants to accept this Code of Conduct. Reporting: if someone makes you or anyone else feel unsafe or unwelcome, or if you believe a harassment problem exists, please report it as soon as possible to Anna van 't Veer (email: a.e.van.t.veer@fsw.leidenuniv.nl) or any of the other organizing team members, either in person or anonymously.

Contact us

If you have any questions, please contact one of the organizers:

Noah van Dongen
University of Amsterdam
n.n.n.vandongen@uva.nl

Daniel Lakens
Eindhoven University of Technology
d.lakens@tue.nl

Anne Scheel
Utrecht University
a.m.scheel@uu.nl

Felipe Romero
University of Groningen
c.f.romero@rug.nl

Anna van 't Veer
Leiden University
a.e.van.t.veer@fsw.leidenuniv.nl

Acknowledgments

Thanks to Marianne de Bruin from Eurandom for the organizational support.

Abstracts (alphabetical order)

Ana Martinovici

Is this real data? How to detect data fabrication in Qualtrics questionnaires

Qualtrics is one of the most used platforms for online data collection. Due to having relatively easy-to-use point-and-click functionalities Qualtrics is used by both researchers and students. Some of these functionalities allow the owner of a Qualtrics survey project to modify previously collected answers and thus engage in actions that raise suspicions of data fabrication or falsification – two of "the clearest examples of research misconduct" (Netherlands Code of Conduct for Research Integrity, 2018, Chapter 5, section 5.2.A.1). Another way to fabricate data in Qualtrics studies is by taking the survey multiple times while claiming to be a different respondent. In this talk, I will show how to check for signs of (1) changes of previously collected answers, and (2) repeated entries from a single respondent. These checks are informed by my experience discovering and reporting suspicions of data fabrication in theses written by students (MSc).

There is no reason to believe that only the students I supervise are engaging in this form of misconduct. Thus, it is possible that data fabrication and falsification is taking place much more frequently than the number of cases we currently detect – both in work done by students and research published in academic journals. This has implications for the design and implementation of open data policies, and for procedures that verify compliance with these policies.

Andrea Kis, Elena Mas Tur, Daniël Lakens, Krist Vaesen, Wybo Houkes

Leaving academia: PhD attrition and unhealthy research environments

This study investigates PhD candidates' (N = 391) perceptions about their research environment at a Dutch university in terms of the research climate, (un)ethical supervisory practices, and questionable research practices. We assessed whether their perceptions are related to career considerations. We gathered quantitative self-report estimations of the perceptions of PhD candidates using an online survey tool and then conducted descriptive and within-subject correlation analysis of the results. While most PhD candidates experience fair evaluation processes, openness, integrity, trust, and freedom in their research climate, many report lack of time and support, insufficient supervision, and witness questionable research practices. Results based on Spearman correlations indicate that those who experience a less healthy research environment (including experiences with unethical supervision, questionable practices, and barriers to responsible research), more often consider leaving academia and their current PhD position. In my talk/poster I would

like to present these results, add recommendations based on our data and the literature, and outline connected lines of research.

Aurélien Allard

Theory building and replicability: on the value of basic facts without theoretical foundations

The last 10 years have seen increased attention paid to the reproducibility of scientific experiments. This focus on reproducibility has met some pushback. One particularly interesting backlash from a philosophical point of view has concerned a group of psychologists and cognitive scientists who have promoted the superiority of theoretical concerns over pure replicability concerns (Buzbas & Devezer, 2023; Devezer et al., 2019, 2021; Feest, 2023; Flis, 2022; Haig, 2022; van Rooij & Baggio, 2020). According to these theory proponents, the focus on replicability is misguided, and risks to backfire if it is not supplemented or replaced with increased attention towards theory building.

Theory-reformers have put forward two main arguments in favor of focusing on theory-building, rather than on promoting reproducibility. The first argument is based on the premise that the value of replicability is dependent on theoretical sophistication. According to this idea, replicability without theory is of little value (Buzbas & Devezer, 2023; Devezer et al., 2021; Muthukrishna & Henrich, 2019). One major reason behind the no-value-without-theory argument relies on the idea that it is hard or even impossible to interpret experimental results without the underpinning of a proper theory, since theories are necessary to identify experimental effects. I call this idea the identification argument.

The no-value-without-theory argument is reinforced by the theory-as-means-towards-replicability argument (Muthukrishna & Henrich, 2019). According to this second argument, focusing on replicability here and now would be a bad idea even if replicability was our sole aim. In this framework, improved theory can be seen as both the ultimate goal of science, and as a means towards replicability. This second argument takes place within a global opportunity cost argument: there are more urgent issues to address than direct replicability issues, and scientists should prioritize other areas than purely methodological concerns.

This talk includes both a positive and negative contribution. On the negative side, I show the limits of the arguments promoted by theory reformers. On the positive side, I provide a general framework to understand the link between theory-building and the promotion of reproducibility.

I begin by examining the normative value of both building theory and establishing a-theoretical facts. Due to virtues such as simplicity, applicability, and breadth, I defend the major importance of theory building in science. However, this general high value of theories does not preclude the importance of establishing basic facts, especially if these facts contradict common beliefs among the scientific community.

Second, I examine the identification argument, and find it lacking. While some underlying assumptions are indeed necessary for inference, psychologists seem to be able to identify factors that are useful to identify experiments and effects.

Third, I examine the opportunity cost argument, and argue that, contrary to the assumptions of theory-reformers, opportunity costs generally favor establishing reproducibility practices as a means of promoting theory-building. While theory building can in some contexts contribute to reproducibility, this influence is not strong enough that it should preclude efforts at improving reproducibility on its own. Overall, improving reproducibility and improving theory should be seen as mutually reinforcing.

Auste Valinciute

Correction of scientific errors in the media

Retractions are an important tool for correcting the scientific record and curbing the circulation of scientific errors and false claims in the academic literature. However, scientific errors and false claims sometimes spread beyond the academic community before they are retracted, for example, via the public media.

This talk will address the challenges that increasing rates of retraction pose to public science communication. It will present results of a mixed-methods content analysis of media mentions featuring retracted scientific publications on COVID-19 (n=945). This content analysis explored if and how media platforms correct news stories in which the retracted papers previously appeared, and how these papers were initially covered.

Results showed that certain initial media mentions of publications that were later retracted featured some levels of criticism (23%). Some media platforms (8%) update or edit news stories, once the scientific publications they feature are retracted, but correction practices are still rare. Media platforms that correct published news stories are mostly the known mainstream news organizations and their affiliates, and popular specialty sites. Yet even among these media platforms, the practice is not consistent. The most common presentation of corrections are top-line statements, notifying readers that the featured scientific publication has been retracted. Media platforms usually inform readers why a retraction took place, but don't always provide clear explanations how exactly the retraction influences the claims presented in the news article.

This talk will highlight the need for science communication practices that enable the public to understand the causes, meaning and implications of retractions in academic literature, as well as the broader need to (re)think about discourses around correction of science that support public understanding and trust.

Bente Sinke, Matteo Colombo, Michal Klincewicz

Scientific credit and the Matthew effect in neuroscience

According to the Matthew effect, scientists who have previously been rewarded are more likely to be rewarded again. Although widely discussed, it remains contentious what explains this effect and whether it is unfair. Three factors relevant to clarifying these issues are examined: scientists' fecundity in supervision, H-index as measure of academic success quantifying output, and the location where their PhD was awarded. This study aims to clarify such relationships, relying on NeuroTree (https://neurotree.org), a large crowdsourced online genealogy documenting the information of neuroscientists, focusing on PhD mentor-mentee relationships. We find an association between location and H-index, but no association between fecundity and H-index. While fecundity in supervising many PhD students might therefore not explain the Matthew effect in neuroscience; geographical location seems a more plausible factor thus entrenching unfair status hierarchies in the scientific credit system not because of exploitative supervisors but partly because of lucky geographical factors.

Exploring existing differences in regional or local patterns, high values of H-index appear to be concentrated most prominently in Japan, the Eastern coast of the USA, and Western Europe, signifying 'hotspots' where neuroscience is being practiced. These clusters are relatively centralized, compared to the low H-index scores which appear more dispersed across the globe. Contextualizing these results, the high-to-low clusters appearing similarly dispersed, could be considered to represent outlier academics with high H-index scores living in remote areas where their general surroundings do not match their exemplary academic credit. The low-to-high clusters appearing once again concentrated in the USA and Western Europe, could be interpreted as resembling the hierarchal structure of academia in which junior scientists are drawn to senior scientists, seeking to collaborate and trying to advance their academic careers. Over time, these junior scientists accumulate credit themselves and attract those with lower H-index scores to engage in collaboration, starting a new cycle of low-to-high clusters or becoming a high-to-low cluster.

These findings contribute to a more nuanced interpretation of the Matthew effect by undermining the idea of a genius single-handedly moving the field and society; as well as singling out mentorship as one of the most salient contributions to society. Instead, science is increasingly done by communities, functioning differently depending on location. These scientists may be attracted to specific locations for other reasons, however, such as higher salary, better infrastructure, or quality of life, these findings make it salient how luck plays an important role in explaining the allocation of credit in science. After all, the geographical location of one's PhD is associated not only with reputation, but also with differential availability of financial resources, equipment and infrastructure and such material differences have less to do with researchers' competence or merit than with historically lucky, social, political, and economic processes.

Candida Sánchez Burmester

Travelling claims: changing presentations of a nanoparticle in different genres

There have been growing concerns about scientific errors in the field of nanotechnology. Various sloths have pointed out the fabrication and manipulation of data in this field. In many of these situations the distinction between correct and incorrect information is clear-cut. However, in other cases there seems to be a grey area where it is not entirely clear if something counts as an error.

In my talk I will focus on a case that falls into this grey area. I will follow travelling claims about a specific nanoparticle and analyse how claims about this particle change as they move from genre to genre. The particle of interest is called spherical nucleic acid, yet, it is telling that even its name changes. I will present how statements about properties and in some cases also about the nature of this nanoparticle differ across genres.

My analysis starts in 2006 when the particle was suggested for controlling protein expressions in cells, and ends in 2021 when a fraud case of misreported data on a preclinical program was reported in one of the companies developing this particle. Considering that this nanoparticle was meant to be promoted both as a commercial product and pharmaceutical drug, I have included different types of genres geared at researchers from academia and industry (e.g., scientific articles, scientific newspaper articles, patent applications, legal company documents, conference proceedings, award speeches), as well as genres that are meant to address groups in the broader public (e.g., public newspaper articles, press releases, blog posts, Wikipedia articles, tweets, publicly recorded talks, and interviews).

I will argue that increased hedging and over-simplification can lead to misinformation which, in extreme cases, can result in errors. Some claims about spherical nucleic acids presented in scientific articles include moderate hedging. In a scientific article in 2007 it is, for example, indicated that this nanoparticle is “potentially very useful” since it can enter cells and detect and regulate RNA. Following this particular claim across different genres shows that the hedging increases and leads to over-simplification. For example, in a patent application filed in 2013, the particles ability to detect RNA is presented as being useful for treating a large range of diseases, including infectious diseases, allergies, autoimmune diseases, and cancer. This and other examples illustrate that the presented (un)certainty of this particle changes depending on the outlet. Even though the examples are taken from nanotechnology, I suggest that differences in the ways claims are presented across genres are a phenomenon that could also be observed in other disciplines and fields. Becoming aware of claims that change across genres could help to identify misinformation and prevent errors from developing.

Catarina Dutilh-Novaes

Keynote: Mistakes in mathematical proofs

Imre Lakatos famously claimed that mathematical knowledge is produced by a dialectic of 'proofs and refutations', whereby a proof-concept is proposed which is then scrutinized to ascertain whether there are counterexamples to individual steps in the proof (local counterexamples) or else to the whole proof (global counterexamples). In my book The Dialogical Roots of Deduction, I further develop this insight in terms of Prover-Skeptic dialogue where Prover tries to prove a conclusion from given premises, while Skeptic critically examines the proof not only looking for counterexamples but also for steps in the proof that are not sufficiently clear.

I submit that the Prover-Skeptic model provides a compelling account of practices of mathematical proof in real-life mathematics. In this talk, I present the Prover-Skeptic model and show how it illuminates practices of proof in mathematics, in particular peer review and how proofs are certified within the relevant mathematical community. I then discuss three examples of Prover-Skeptic interaction in mathematical research: Wiles' proof of Fermat's last theorem, a failed proof of the inconsistency of Peano Arithmetic, and a purported proof of the ABC conjecture whose status as a valid or invalid proof has been under debate for over 10 years now.

Christian Hennig

Understanding statistical inference based on models that aren't true

Statistical inference is based on probability models, and most of the theory behind it assumes these models to be true. But models are idealisations, and it makes little sense to postulate that they are literally true in reality. Models are however required to analyse the behaviour of statistical methods in any generality. In order to explore the implications of running statistical inference based on models that aren't true, it is helpful to look at more general supermodels that allow for violation of the supposedly assumed models. I will present a framework for how to think about statistical inference based on models that aren't true, conditions under which such inference can be useful or misleading, and what impact this has on the interpretation of the results in practical settings.

Cristian Mesquida

What is your research hypothesis? On the importance of deconstructing your research hypothesis to improve the severity of hypothesis tests

In response to the exploitation of researchers' degrees of freedom, most recommendations for improving methodological rigour focus on methods and procedures that scientists use

to collect, analyse and publish data. However, a higher-level and unsolved issue is how research hypotheses are stated, and then translated into a statistical hypothesis. In the Neyman-Pearson approach to null hypothesis significance testing, hypothesis tests cannot be used as direct tests to test the research hypothesis. Instead, researchers must translate their research hypothesis into a statistical hypothesis, which is expressed as a pair of complementary parameters: the null hypothesis and the alternative hypothesis. Once the two statistical hypotheses have been set up, a hypothesis test is conducted as an indirect test and the calculated p-value is then used for the “rejection or not-rejection” of the null hypothesis, which leads to the “acceptance or not-acceptance” of the alternative hypothesis resulting in the “support or not-support” of the research hypothesis. This requires a logical derivation chain whereby the research hypothesis should be a (quasi)identical statement to its statistical hypothesis. Otherwise, any mismatch between the research hypothesis and its statistical hypothesis reduces the severity of hypothesis tests making it easier for researchers to find support for their research hypotheses. Inspired by the seminal paper of Hand (1994) “Deconstructing Statistical Questions'', we deconstruct a series of research into basic statements and then use truth tables to determine whether the research hypothesis and its statistical hypothesis are logically equivalent and therefore, (quasi)identical. The aim of this work is to stimulate debate about the need to formulate research hypotheses sufficiently precisely that can be unambiguously and correctly matched with their statistical hypotheses.

David F. Urschler

Do we really know what we are talking about? The prevalence of well- and ill-defined psychological constructs in 150 meta-analytic reviews.

The importance of scientific rigor has been in the spotlight of psychological research in the past decade. For example, previous research has provided comprehensive guidelines to increase replicability, and highlighted the importance of theory for improving psychological science. However, we argue that the crucial aspect of construct clarity has not yet been sufficiently considered, because concept clarification of psychological constructs is a prerequisite for reliable and valid scientific endeavors. Our reasoning is underpinned by previous research that has revealed that several eminent psychological constructs have been ill-defined. For example, a review revealed that empathy has been defined in 43 different ways, which has had a negative impact on both research and practice. Moreover, constructs of interest have been solely defined by their operationalizations (i.e., operational definitions) that undermines conceptual clarity. Consequently, we examined the prevalence of well- and ill-defined constructs in psychological science.

Given that meta-analytic reviews on a certain topic are more likely to be cited and to influence the field than a single research paper, we sampled 150 meta-analyses from Psychological Bulletin across the period from 1990 to 2017. We focused on Psychological

Bulletin because Psychological Bulletin is a renowned outlet for meta-analytic reviews across all psychological disciplines. To answer whether the construct(s) of interest were defined, and if yes, how were they defined, each meta-analytic review was coded by two independent coders. Additionally, we coded several descriptive characteristics of each meta-analytic review (e.g., number of included constructs, discipline, first- and last-authors gender, construct's valence; the full-list of coded characteristics will be available at OSF).

Our results revealed that the 150 meta-analytic reviews contained 359 constructs in total. Out of these 359 constructs, 140 (39%) constructs were defined. Out of the 140 defined constructs, for seven a conceptual definition, for 50 a homeostatic property cluster definition (a construct is defined by a cluster of features that regularly but not exceptionlessly co-occur), and for 83 an operational definition was provided. The main findings are, that the majority of constructs in meta-analytic reviews were not defined, and if they were defined, the majority of the constructs were lacking a conceptional definition. In our discussion, we argue for an increased emphasis on conceptual definitions, which, in turn, further improves scientific rigor in psychological research in the future.

Dimitri Paisios, Nathalie Huet & Elodie Labeye

The Dark Side of Likert-type Scales: Implications of the Midscale Disagreement Problem

Proper stimulus control in psychology experiments plays a fundamental role in the validity of their results. In many cases, the criteria used for stimulus selection include dimensions (e.g. concreteness, emotional valence/arousal) assessed behaviourally through Likert-type scales. Typically, a pilot group of participants is asked to rate a list of potential stimuli on a scale. The average rating for each item is then computed and used to determine which items fit the criteria to establish the experiment's stimulus lists. Recently, several factors such as technological advances, the need for standardised materials and a high number of theoretically relevant dimensions to control have also led to a proliferation of large rating databases which provide standard summary statistics for hundreds to tens of thousands of items. Despite their importance and the increasing amount of resources dedicated to their collection, however, the ratings in themselves have been subject to surprisingly little methodological consideration. One of the key assumptions in the aforementioned approach is that the average rating reflects the item's position on the scale's continuum. Through a case study in psycholinguistics, we show instead that most items with an average rating towards the middle of the scale display high disagreement among raters, and thus that their averages do not, by themselves, capture any meaningful information about the underlying responses. Rather, they are an artifact of the scale. After providing an intuitive graphical interpretation of Likert-type summary statistics (the typically reported means and standard deviations), we derive two additional implications of this midscale disagreement problem and argue that it greatly affects the validity of a large number of studies – either because of inadequate stimulus sampling or statistical modelling. We finally

extend our analysis to some fields in which Likert-type ratings are treated as the dependant variable and show that they also suffer from an inadequate interpretation of their results. We conclude by opening several avenues for research on Likert-type scales and human assessments, as well as a discussion about the importance – and difficulties – of familiarising oneself with the raw data before undertaking any statistical procedures.

Fiona Fidler

Can forecasts of replicability improve peer review?

Now in its sixth year, the repliCATS project (Collaborative Assessment for Trustworthy Science) has evaluated over 4,000 published social science articles across 8 disciplines, including psychology, economics, and education, as well as many preprints. For each paper, a diverse group of experts forecasts the likely replicability of the research findings and makes a variety of other judgements about the credibility of the evidence presented using a structured deliberation protocol. This talk will present our approach to evaluating research, and for cases where we have the outcome of actual replication studies, data about the accuracy of our forecasts. I will also discuss how structured expert elicitation, deliberation, and decision protocols like those used in repliCATS might improve peer review more generally.

Gerit Pfuhl, Adrian Helgå Vestøl, Ole Fredrik Borgundvåg Berg

Do Norwegian publishing incentives lead to questionable research practices?

The way from data to publishing is distorted by incentives. In Scandinavia journals are classified into levels, and the higher the level the more points a publication earns. At the same time, these "better" journals are often journals that publish novel and significant results. We therefore hypothesized that questionable research practices might be more prevalent in higher level journals than in lower level journals. We collected data over the last 20 years from researchers working at psychology departments at the four major Norwegian universities. We extracted the statistical results and will perform the z-curve analysis (observed discovery rate, expected discovery rate and expected replication rate). We look forward to present the results at the conference.

Ian Hussey

Keynote: The ERRORR project: A three-prong effort to improve post publication critique and error detection.

By itself and as currently implemented, academic peer review is not up to the task of comprehensively detecting errors in scientific publications, at least in the field of

psychology. Error detection typically requires more resources than available for peer review: there are simply too many manuscripts, and errors can come in too many forms. Post publication scrutiny and critique, for example of influential or controversial claims, represents an important parallel system of scientific verification. Unfortunately, published work is rarely checked for errors, likely because this behavior is poorly rewarded, there is a shortage of relevant tools, and little training in error detection is available. In this talk, I will discuss our efforts to improve each of these issues through the ERЯOR project (Estimating the Reliability and Reproducibility of Research). First, borrowing the concept from cybersecurity research, the ERЯOR project is the first large scale Bug Bounty program for psychological science. Published research findings are scrutinized for errors, with monetary payouts to the authors vs. the error checkers contingent on whether errors are found. Second, the ERROR project will produce resources and training materials. Existing tools are being collated and documented and an R package is currently under construction. Lastly, I will discuss the masters' degree course in post publication error detection I teach, and the need for comparable courses at other institutions – echoing Dorothy Bishop's recent call in her blog post "Defence against the dark arts: a proposal for a new MSc course".

Jamie Cummins

Erring on the side of caution: Family-wise Type II error

In the null hypothesis statistical testing framework (NHST), the concept of the Type I family-wise error rate has been extensively studied and discussed. When (and when not) to correct alpha for multiple comparisons, and the methods by which to do so (e.g., Holm-Bonferroni correction, Šidák correction) are well-established. However, the Type II family-wise error rate has not received comparable attention. This talk will discuss the concept of Type II family-wise error and clarify its relationship with Type I family-wise error. Using simulations, I will demonstrate the impact of multiple-testing on Type II error/statistical power, highlight when (and when not) family-wise Type II error is relevant to consider, and discuss how best to balance between Type I and Type II error control within families of tests.

Judith ter Schure

Keynote: Making research Evidence-Based ruins error control from the sampling space

In 2009, a 'radical' idea was proposed to reduce research waste in biomedical sciences: informing new research by past results. This recommendation has been reiterated in every research waste paper since, and embraced by the movement to do Evidence-Based

Research. Surprisingly, this proposal is actually radical, when we care about (frequentist) error control and coverage of intervals, that guarantee properties averaged over the sample space. The reason is that we cannot define the sample space of two studies – let's say two randomized clinical trials – if the design of the second study is informed by the first. If depending on the findings in the first, the population definition (e.g. inclusion criteria) is decided, the outcome measure might be defined in a new way, or a second trial is unethical because the first showed treatment harm. This talk will discuss accumulation bias: if the existence of future studies depend on the results of earlier ones in the same meta-analysis, or its timing might be changed. It will also discuss the need for new foundations of statistics: if the sample space is so flexible that it cannot even be described by standard probability (measure) theory. And it will discuss so-called anytime-valid statistics based on e-values and foundations of probability and statistics based on betting, that can resolve both.

Leonhard Volz, Denny Borsboom, Noah van Dongen

PsychoModels: Giving strong theory building a FAIR chance

Formal theories and their computational implementation have been widely suggested as an antidote to the reproducibility crisis. A formal description of psychological processes offers a precise formulation of mechanisms and relations, allows clear communication about assumptions, and enables deducing and testing implied hypotheses. However, formal modelling is not an established practice in the wider psychological research community and much of an enigma to most researching psychologists, as well as a quite idiographic enterprise for individuals who apply modelling to their work. To this end, we started a research project that, at its core, aims at creating a database that facilitates sharing, (re)using, and extending computational implementations of formal models of psychological processes.

This curated collection ensures that included models are clearly annotated and presented in a similar way, making it easier to skim a model and grasp its content than currently existing repositories allow for. Furthermore, the similarity in presentation facilitates a common language around modelling that facilitates communication between researchers. Also, indexing model characteristics allows for search options across the model database, facilitate model review efforts, and promotes reuse of the models in the database. Lastly, easing access to existing models and connecting educational materials to the database will allow researchers with less experience in formalising their research to learn from best-practice examples.

Our talk outlines our proposal for an indexing system that captures relevant aspects of (implementations of) computational models of psychological processes, and presents a

work-in-progress version of a platform that implements our taxonomy and additional functionality. We expect that our efforts will benefit theoretical and empirical work, aid didactic efforts around modelling, and ease access to computational modelling literature and materials. We hope that this presentation can spark a discussion about how to facilitate formal psychological research and practical steps to support a community of computational modellers.

Ligaya Breemer, A. E. van 't Veer, L., P. M. Isager, T. Heyman, T. van Leeuwen, & M. C. Makel

The Prevalence of Replications in Psychology Revisited

Numerous reforms have taken place in the last decade, with an increasing awareness that replication is crucial to the scientific process of knowledge creation and revision. In 2012, replication prevalence in psychology up to that year was estimated at 1.07% by Makel, Pucker and Hegarty. Conceptually replicating their work, we estimated replication prevalence in the last decade (2011-2020) by coding a random sample of a thousand empirical psychology articles that use the word replicat*. We discuss the results along several themes: replication type (direct or conceptual), successfulness, authorship overlap, subfield, publication year, and bibliometric analyses on the citation, journal, and author level. Results reveal a continuation of the upward trend found in previous research, with a newly estimated replication prevalence of 2.3%. The replication prevalence varied somewhat between subfields. Additionally, we found fewer successful and more mixed-outcome replications, and success associated with overlap between the original and replicating author teams. More than half of the replications were considered conceptual replications. The original studies had a higher average citation impact than the replications. Interestingly, failed replications generally had a lower citation impact, but received more social media attention.

Olmo van den Akker, Marcel van Assen, Marjan Bakker, & Jelte Wicherts

Preregistration: Past, Present, and Prospects

While preregistration has been lauded as one of the solutions to the replication crisis in psychology, not much empirical evidence is available about its effectiveness. In this set of studies, we aimed to assess whether preregistrations in psychology are sufficiently producible (i.e., they can be conducted based on the information provided in the preregistration) and sufficiently in line with the corresponding publications. We also assessed whether preregistered studies include a lower proportion of positive results than non-preregistered studies, which would be an indication of a preventative effect on questionable research practices like p-hacking and HARKing. We assessed 459

preregistered studies that either won a Preregistration Challenge prize or earned a Preregistration Badge. We custom-made checklists to assess preregistration producibility and preregstration-study consistency. More than 30 coders used these checklists to assess the studies in our sample. We selected our control group of non-preregistered studies based on the 'related records' function of Web of Science. We found that there is room for improvement for preregistration in psychology. Hypotheses, statistical models, and inference criteria were typically not very well described in preregistrations. Moreover, we found that the consistency between preregistrations and papers was low, mainly for data collection procedures and statistical models. More comprehensive preregstration templates did lead to more producible preregistrations. When comparing preregistered and non-preregistered studies we found no difference in the proportion of positive results, but preregistered studies were typically of higher quality and had more impact than non-preregistered studies. There is room for improvement with regard to the effectiveness of preregistration in the field of psychology. Although it could be that preregistrations, especially when they are based on a comprehensive template, prevent some questionable research practices, the practice of registered reports may be more promising.

Peter Stilwell, Tom Heyman

Sequential Strategies: Assessing Group Sequential Designs with Real Data

This study embarks on an exploratory investigation into the feasibility of group sequential testing approaches within psychological science, contrasting them with the traditional fixed sample size approach. The research is grounded in the recognition that fixed sample size approaches often lead to suboptimal data collection; either insufficient or unnecessarily excessive. Our aim is to critically examine whether group sequential testing, a dynamic method where sample size is adjusted based on interim analysis outcomes, could offer a more efficient alternative without compromising the integrity of research findings.

The study is structured around three principal research questions: 1) the potential efficiency of group sequential testing in reducing sample size compared to the fixed N approach, 2) its influence on the outcomes of hypothesis testing, and 3) its impact on the accuracy of parameter estimates. We employ a retrospective application of sequential analyses on data from nine multilab registered reports, allowing us to assess the method across various datasets and experimental conditions. This approach enables an evaluation of sequential testing in real-world scenarios, rather than relying solely on theoretical or simulated data. The study design includes variations in the number of interim analyses and considers both efficacy and futility as criteria for terminating data collection.

Through this investigation, we aim to provide empirical insights into the practicality of group sequential testing in psychological research. The outcomes of this study will

contribute to the ongoing discourse on optimal research methodologies in the field, particularly in relation to efficiency, reliability, and validity of research conclusions.

Raphael Merz, Stephen Lee Murphy, Linda-Elisabeth Reimann, Aurelio Fernandez Zapico

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), and whether evidence support the hypothetical possibility that researchers generally mean an effect is likely too small to matter when they make no effect statements. Accordingly, the present study aims to investigate these points – to clarify the prevalence of nonsignificance misinterpretations in published psychology articles, see examine whether this issue improved over the past decade, and to explore whether researchers generally know that nonsignificance does not reflect an effects absence. To achieve these aims, we highlight nonsignificance statements in the discussion sections of 600 articles across three time-points (2009, 2015, 2021) from ten psychology journals of varying impact factor that contained a nonsignificant finding. We then code each statement as correctly or incorrectly interpretating nonsignificance, and whether incorrect interpretations were sample-focused (e.g., ‘age did not affect our dv’) or population-focused (e.g., ‘age does not affect self-control'). We also code the article publication year and extract effect sizes and textual interpretations of a misinterpreted nonsignificant effect and a significant effect contained in the same article. We aim to present our findings at the conference in late February. Relatedly, we will also reveal the progress we have made in leveraging AI to take over, accelerate, and simplify the ‘hard-coding’ of academic manuscripts, and our ideas on how AI algorithms, when trained, tested, and refined using extant hard-coded articles, may supercharge meta-scientific progress. We also present how such technological developments could enable the creation of a ‘statistical spellchecker’, that provides users with accurate feedback on their statistical interpretations and more general article content (e.g., the absence of any data availability statement).

Rink Hoekstra, Nina Schwarzbach, Henk Kiers, James Steele, & Fiona Fidler

The Illusion of Objectivity - Discussing the Hidden Subjectivity in Quantitative Social Science

In the social sciences, subjectivity is often considered as something that preferably needs to be prevented, while objectivity is seen as an honorable objective. This may explain why quantitative research seems to be considered of higher scientific status than qualitative research, in which subjectivity has a more explicit role. We argue that, regardless of the type of method used, the question should not be whether we should be subjective in our work, but how to communicate (about) it. We will 1) discuss the perception of subjectivity and objectivity, 2) call to rewrite the narrative of objectivity as a scientific virtue, and 3) address what this entails for different stakeholders inside and outside science. Thus, we intend to highlight the significance of reflecting on and accepting subjectivity in conducting, perceiving and consuming research, to eventually enhance transparency and promote a deeper understanding of the inherent complexities within the social sciences.

Simine Vazire

Keynote: The dog that caught the car: Turning scientific values and principles into journal policy and practice

Having spent many years working on improving psychology's research and publication practices, I was recently given a unique opportunity to try to put my talk into action, as Editor in Chief of a large and well-respected journal in psychology (Psychological Science). Psychology's recent reform efforts have been driven by values and principles that can sometimes be quite abstract. Moreover, it can be hard to find real-world contexts in which to test or implement reforms at scale, making it difficult to anticipate obstacles or constraints. In this talk I will reflect on my experience over the last few months of working with the editorial team at Psychological Science, and with the broader community, to move closer to scientific ideals in publishing and peer review. Some steps have been obvious, easy, and (I think) uncontroversial. In other cases, this experience brought to light pragmatic constraints, competing values, or other considerations I had not anticipated. As is common in implementation science, the experience of putting ideals into practice provides fertile ground for reflecting on our goals and approaches in scientific reform.

Taym Alsalti

Meta-analysis: Is There Still Hope?

Meta-analyses are often characterised as being at the top of the evidence hierarchy: they are cited more frequently than primary studies about the same topics, are commonly assumed to provide the most accurate estimate of an effect, and have a considerable impact on theory development as well as policy and clinical practice. Recognising their importance, experts have, over the last 3 decades, developed several guidelines for planning, conducting, analysing, and reporting the results of meta-analyses. At the same time, meta-analysis is anecdotally often described as a "quick and easy" publication. This might explain the meta-research showing that even relatively undemanding guidelines like PRISMA are rarely fully adhered to, when at all. Reviews evaluating meta-analyses in more detail across many fields have shown that meta-analyses are often irreproducible, heedless of obvious sources of bias (e.g., publication bias), and generally fraught with methodological and conceptual errors. So, on the one hand, everyone interested in conducting a meta-analysis has access to a wide range of resources for doing so rigorously and transparently, on the other hand the behavioural, social, and medical sciences are filled with faulty meta-analyses that seem to have been the product of wilful or negligent ignorance of these resources. What, then, do we do with all these faulty meta-analyses? Is an irreproducible/insufficiently reported/methodologically unsound meta-analysis useless, or even harmful if used to inform research, clinical practice, or policy making? Given that historically it was not easy to convince journals to convince journals to implement more stringent editorial procedures, how can we improve the quality of future meta-analyses? I will use this poster to present results from my work and others' relating to the patterns mentioned above and discuss potential answers to the larger questions posed here.

Veronika Cheplygina, Amelia Jiménez-Sánchez, Gaël Varoquaux

Shortcuts and other shortcomings in machine learning for medical imaging

The application of machine learning (ML) to medical imaging diagnosis has attracted a lot of attention in recent years, with numerous reports of recognising medical images more accurately than human experts (for an overview see Liu et al., 2019). Yet progress in clinical practice has not been proportional to claims. For example Roberts et al. (2021) found that none of the 62 published studies on ML for COVID-19 had potential for clinical use. Studies for other clinical applications of ML have also failed to find reliable published prediction models.

The increased popularity of ML in recent years is often explained by two developments. First, there are several large publicly available datasets. Second, open source deep-learning toolboxes allow development of algorithms without specialised domain knowledge, allowing more researchers into a field. Despite these seemingly ideal conditions for reproducibility, the state of ML in medical imaging is not as positive as one might think. There are various reasons for this which we outline in (Varoquaux and Cheplygina, 2022), here we highlight two.

One reason is that large sample sizes are not a panacea. There is a tendency to expect that a clinical task can be "solved" if the dataset is large enough. However, not all clinical tasks translate neatly into ML tasks. Furthermore, creating larger datasets often comes at the expense of quality, leading algorithms to learn spurious correlations or "shortcuts". For example, an algorithm might learn that if a patient's chest x-ray shows a drain – a treatment for a collapsed lung – that that patient is likely to suffer from the collapsed lung condition (Oakden-Rayner, 2020). Similarly, our recent results (in preparation) show that lung diseases can be diagnosed with high accuracy, even if the lungs are hidden from the x-ray.

One reason is that the availability of data and code, plus the theoretical option to "infinitely" repeat experiments (for example, with different subsets of data, different initialization points of the algorithms, and so forth) creates an illusion of generalization. Since there are many degrees of freedom to how such repetition can be done, for practical reasons researchers tend to not do this exhaustively, but might be tempted to formulate their conclusions more generally.

In this talk I dive deeper into these problems and hopefully, with the help of the audience, also explore some solutions.

Hackathon Abstracts

Hackathon #1: Risk of Statistical Error (ROSE)

Rickard Carlsson & Natalie Hyltse

In systematic reviews with meta-analyses, researchers routinely conduct quality assessments of the included primary studies. A common approach is to assess the studies' Risk of Bias (ROB) based on a checklist, labeling them as low risk, moderate risk, or high risk of bias. Currently, these tools only assess the methodological quality of the studies and do not consider if there are errors present in their statistical reporting and/or dataset that might bias the result (beyond bias due to questionable research practices).

We suggest the development of a tool to detect/predict the presence of error in primary studies: Risk of Statistical Error (ROSE). Akin to other similar assessment tools, ROSE will mainly consist of a checklist (incl. both novel items and well-established checks/tools). The focus of this new checklist is only to detect the risk of errors of any and all kinds, ranging from fabricated datasets to simple reporting errors. In other words, we do not aim to develop a tool for determining the cause for errors (e.g., honest mistakes, software malfunction, carelessness, or questionable scientific conduct). Further, similar to ROB, the aim is not to prove the presence of errors, but merely to assess studies on the risk that the data might not be correct.

After an initial on-the-fly unconference at SIPS 2023, we are still in the early conceptualization stages and seeking peer input and discussions. During this hackathon, we present our preliminary work on developing ROSE and invite you to join us in narrowing down areas and items to include in the first version of the checklist. Our goal is to reach a draft of the checklist at the end of the session that can proceed to user testing.

Hackathon #2: Automating checks for reporting standards, statistical inferences, and open science practices

Daniël Lakens

Best practices continuously improve, and due to time constraints it can be difficult for scientists to keep up with developments in the scientific literature. Accessible tutorials and reporting guidelines aim to make it easier for researchers to adopt best practices, but even these are not read as widely as they should be. Additionally, researchers might not remember to use best practices when they write their articles. In this hackathon approach this problem from a human factors perspective, and examine where we can automatically detect the absence of best practices in scientific articles researchers write. Similar to a tool like StatCheck, we created a tool in Python that reads in text from a pdf and automatically

identifies relevant content. For example, we check for open science practices, such as whether text contains links to online repositories, and if so, if those repositories have been made publicly accessible (which researchers often forget). We also check for adherence to reporting guidelines, such as whether researchers report exact p-values (instead of $p < .05$). Finally, we explore the possibility of natural language processing algorithms to correctly classify statistical inferences, and attempt to automatically detect sentences in which researchers incorrectly conclude the absence of an effect based on $p > .05$. Our goal in this hackathon is to evaluate the usefulness of the preliminary version of the tool we have created, extend the number of best practices that are automatically screened for, and increase the accuracy of the detection of possible improvements.

We are interested in the following questions:

Which suboptimal practices can be detected through "rule-based" approaches? For example, finding terms like "marginally significant" or "observed power". Note that a rule-based approach can also compute information (as Statcheck does) or follow a link and check the information (such as whether an OSF page is open).

Which suboptimal practices can be detected through natural language processing approaches? For example, classifying statistical inferences about p-values and bayes factors are correct or incorrect.

Which data sources exist where meta-scientists have classified the absence of best practices, and can they be used in rule-based or natural language processing algorithms?

Which meta-scientific questions could we answer with a tool like this?

Notes:
https://docs.google.com/document/d/1l1Z_4cGKlMzgUA9_kdophf5Z92ItQFVjDg0ytqJ34LI/edit?usp=sharing

Rule-based:
https://colab.research.google.com/drive/18LnZf6ZFhoQhr3Bcg4Wrk7Q0zQyGRu2G?usp=sharing

NLP: https://colab.research.google.com/drive/1PM8Ur97vNhFJmb3J0PXkDiXrDAe2T89a

Data generation:
https://colab.research.google.com/drive/1unoig-onH-Z5kVqnuFjbFq82EEQvEQae?usp=sharing

Hackathon #3: Assessing Computational Reproducibility

Lisa DeBruine

Computational reproducibility is the ability to generate the same results with the same data and analysis. While this is a minimum standard for robust research, it is surprisingly difficult to assess. This hackathon will introduce you to the compreprev R package and shiny app for structuring computational reproducibility reviews:https://github.com/debruine/compreprev.

In the session, we will practice using this tool to review open data and code, with the goal of refining the tool and its documentation.

https://github.com/debruine/compreprev

https://rstudio-connect.psy.gla.ac.uk/compreprev/

Hackathon #4: Self-correcting science: Increasing the discoverability and usability of post publication scrutiny and error detection tools

Ian Hussey - landing page

If science is to become genuinely self-correcting, post publication critique will need to be made easier, more normalized, and better rewarded. As it currently stands, we collectively have access to a lot more tools and methods to conduct research than we do to scrutinize published research. This hackathon will contribute to efforts within ERЯOR project to change this by increasing the discoverability and usability of error detection tools.

Contributors could work tasks like the following:

- Gathering existing tools for post publication scrutiny, error detection, and data forensics.
- Documenting their use cases.
- Improving their documentation, vignettes and blogs demonstrating their use
- Brainstorming common error detection use-case that currently lack tools
- Creating to-do lists and roadmaps for existing but under-used tools could be further developed.
- Documenting workflows or creating training materials for scrutiny methods and communication.

People of all skill levels welcome. If you have ever read DataColada, Dorothy Bishop's blog, Nick Brown's blog, or James Heather's work and thought 'we need more of this', this hackathon might be of interest to you.

Hackathon #5: The garage is open – where are the cars? How do we coordinate quality control work in practice?

Peder Isager & Anna van 't Veer - landing page

How can we encourage more organized skepticism? Put differently, how can we organize quality control in a way that is sustainable, and that meaningfully impacts research quality in the long term? The answer to this question has three parts.

First, we need tools for carrying out quality control. On this front we are doing quite well. A plethora of tools and guidelines for error detection and quality control have been developed, and continue to be developed every year.

Second, we need people willing to apply these tools, regularly and consistently. On this front we are doing much less well, but (slowly) progressing. Initiatives such as the Psychological Science Accelerator, journal initiatives such as Psychological Science's STAR editors, and the Dutch NWO replication grant instrument, are examples of organized attempts at giving quality control workers a home base to work from, some incentives for doing the work, and a platform for attracting new talent when existing quality control experts retire.

Third, we need to select the research that should undergo quality control. This is a largely unsolved problem (at least in psychology), and solutions are so far few and far between. "Scrutinize everything" is not a valid solution: quality control costs time and money, which is finite. In practice, all published research cannot be scrutinized. "Build it, and they will come" has largely been our mindset in the metascience community. We hope the research community, once granted access to a quality control mechanism, will know what to do with it. This is not reliably true, and it is not obviously a good coordinating principle. Researchers often do not know what research in their field is important to quality control and have few incentives to find out. When they do know, there are no organized ways for them to communicate their opinions to the quality control workers.

We, the research community, not only need to practice organized skepticism, but also need to find ways to efficiently coordinate collective efforts – what do we agree is important to direct finite resources to? What are the criteria we use to select which cars (what research) get into the garage (the quality control mechanism)? Answering this question is a crucial part of organizing quality control in practice. It is much easier to obtain funding, visibility, and willingness for work that is demonstrably useful to the research community.

In this hackathon we want to discuss how research communities could form an opinion about what research to prioritize for quality control, how best to communicate these opinions to those willing to work on quality control, and how quality control can then be coordinated in practice.