Small Group Meeting

Efficient Science

Methodological Controversies in J/DM Research

26th July 2013

Max Planck Institute for Research on Collective Goods, Bonn

Time table of the workshop

26th of July 2013

Time	Friday
Ab 8:30	Coffee Reception
9:00 – 9:20	Welcome and Introduction
9:20 – 10:20	Keynote: Michael Birnbaum
	Coffee Break
10:35 – 11:35	Keynote: Joseph Simmons
	Coffee Break
11:45 – 12:15	Christoph Engel: Scientific Dishonesty As a Public Bad
12:15 – 12:45	Frank Renkewitz: Random findings, alpha inflation, overrated relevance and generalisability: Consequences of the random variation of true effect sizes.
12:45 – 13:15	Erich H. Witte: Statistical inference techniques: Context of discovery and context of justification in empirical sciences - The long way of a research program
	Lunch Break

14:15 – 14:45	Clintin Davis-Stober: When are our experimental findings better than a guess?.
14:45 – 15:15	Michael Schulte-Mecklenbeck: How replicable are process measures in decision making? The impact of subtle search costs
15:15 – 15:45	Mark J. Brandt: Advancing replicability and theory through replication recipes and replication packages
	Coffee Break
16:15 – 16:35	Christoph Stahl: Data-blind peer review
16:35 – 17:05	Mirjam Jenny: Psychologists are open to change, yet wary of rules
17:05 – 17:35	Andreas Glöckner: The empirical content of theories in judgment and decision making: Shortcomings and remedies
17:35 – 17:55	Klaus Fiedler: How Important is Statistical Hypothesis Testing for the Quality of Science?
	Break
18:00 – 19:00	Discussion
19:00 – 20:00	Dinner

Aim

How can science be done efficiently, so that it serves the collective goal to maximize cumulative knowledge development? Answering this question involves at least two interwoven challenges, one concerning knowledge and another one concerning organizational structure. First, researchers have to know how to do science efficiently, involving aspects of theory formulation, experimental design, methods for data analysis, and theory revision. Second, social dilemma structures of scientific discovery (i.e., mixed-motive dilemmas) have to be overcome in which maximizing self-interest by scientists causes detrimental effects for the scientific community and society overall. These dilemma structures have to be identified and solved and dilemma research can provide guidelines for doing so.

Classic works on theory of science (e.g., Popper, 1934), methodology in psychology (e.g., Platt, 1964), and also on dilemmas (e.g., Dawes, 1980) have addressed these topics. Recent scandals of open scientific fraud, but also findings on the prevalence of unintended minor cheating of "good" people (bounded ethicality, e.g. Chugh, Bazerman, & Banaji, 2005), and the experience of low reproducibility of scientific findings (cf. the replicability project; see also Fuchs, Jenny, & Fiedler, 2012) have brought much attention to methodological issues in the scientific community.

Hence, there seems to be an increased interest in developing methods and standards for efficient science. Some aspects of methodology have received much attention in recent articles (e.g., Bakker, Van Dijk, & Wicherts, 2012; John, Loewenstein, & Prelec, 2012; Nosek., Spies, & Motyl, 2012; Simmons, Nelson, & Simonsohn, 2011; Tressoldi, 2012; Wagenmakers, Wetzels., Borsboom, & van der Maas, 2011), open debates (e.g., opensciencefframework.com, open letter by Kahneman), published special issues (e.g., Glöckner & Hilbig, 2012; Spellman, 2012) and special issues in preparation (e.g., Nosek & Lakens, in preparation; Zeelenberg & Zwaan, in preparation). Although some progress has been made, for example that the problem-awareness within the scientific community has increased and that some journals have reacted to the call for new standards for publication (e.g., publishing data, transparency of procedure and analysis), only a minority of issues have already been solved. Relatively little attention has been given, for example, to more general questions, such as: How can the social dilemma/public good structure of science be solved? How can theory formulation and revision be improved?; And also, what are the potential negative side-effects of changing research practices to focus on replicability and alpha-errors only (e.g., Fiedler, Kutzner, & Krüger, 2012)?

Talks

Scientific Dishonesty As a Public Bad

Christoph Engel
Max Planck Institute for Research on Collective Goods

Much of the debate over scientific dishonesty has a moral undertone. This is understandable, maybe even instrumental. Yet for a moral problem to exist, there should be an incentive to disregard scientific standards. This contribution uses a standard tool from welfare economics, the public bad, to cast additional light on the incentive structure. Arguably, at least in the short run, most scientists would increase their personal utility by being sloppy with scientific standards. Yet if they do it becomes more difficult for all scientists to make their voice heard in society, to convince policymakers to assign public funds to academia, and to lead a fulfilling academic life. Discussing alternative definitions of scientists' utility function helps understand facets of the resulting conflict, and design meaningful interventions.

Random findings, alpha inflation, overrated relevance and generalisability: Consequences of the random variation of true effect sizes.

Frank Renkewitz
University of Erfurt

Much of the current discussion of reproducibility and failed replications seems to rely on the assumption that the true effect size estimated in several direct replication attempts of a finding is constant. In contrast, the recent literature on meta-analysis emphasizes that the appropriate model to analyze several empirical findings on the same topic generally is the random effects model. Meta-analyses in the field of psychology typically confirm this theoretical assumption: The true effect size of primary studies shows variation that cannot be explained by (known) moderators. A widely unappreciated implication of the validity of the random effects model is that significance tests in the primary studies are subject to alpha inflation: When the mean of the true effect sizes is zero more primary studies than expected by the chosen alpha level will find statistically significant results. These significant results are neither theoretically meaningful nor practically relevant. Alpha inflation will occur even if the research is based on a directional hypothesis and the appropriate test is one-tailed.

I will illustrate alpha inflation in studies with randomly varying effects with the example of ten direct replications of a study on fluency and psychological distance (the hypothesis states that cities presented in a less fluent font are judged to be more distant; Alter & Oppenheimer, 2008). In this

case, the pooled effect size of the replication studies is zero, yet four of the studies find a significant effect. Furthermore, I will report results of Monte Carlo simulations investigating how the actual alpha level is affected by the amount of random variation, sample size, and design characteristics of the primary studies. A core result is that even a small degree of random variation may increase alpha by a factor of two or three. Thus, random variation may be a central factor contributing to low reproducibility in psychology. Additionally, the simulations demonstrate that random variation will amplify the effect of publication bias on effect size estimates. Based on the simulations, I will discuss the conditions under which single studies have to remain uninformative and direct replications and meta-analyses are necessary to gain and accumulate knowledge.

Statistical inference techniques: Context of discovery and context of justification in empirical sciences - The long way of a research program.

Erich H. Witte
University of Hamburg

Most of the shortcomings reported in the special sections of the Perspectives (Pashler & Wagemakers, 2012; Spellman, 2012) could be reconstructed by the differentiation of the two classical contexts of research (discovery and justification or confirmation). A whole well-guided research program (Lakatos, 1978) has to follow the long way and not to stop at the context of discovery. The inference methods of the long way are sketched.

The statistical methods have taught us that something which looks clearly systematic might be random. We only can talk about a discovery if we are to some extend sure that the observed results deviate from chance. A single discovery itself is not very convincing in the development of an empirical science. We need several discoveries. The next step then is an integration of these replicated discoveries into a meta-analysis with the known problems of estimating the theoretical parameters (Chan & Arwey, 2012).

After a phenomenon has been discovered empirically it demands a well-founded justification by theoretical modeling. The consequence is testing hypotheses against each other and not against randomness.

The combination of empirical results under the context of justification is the addition of the log-likelihoods determined by the two (or more) separate observational conditions. “No decision” is no more a missing discovery and therefore nearly not publishable, it is a part of the decision process under a well-guided research program in the context of justification (confirmation).

If we accept the idea of a research program with successive proofs of theoretical parameters then we have to find a stopping rule when it is enough for the present to test a theory with its parameters. Such a preliminary criterion might be found by the maximum likelihood Lmax (θi) of the given data compared with the likelihood of the accepted parameter L (θ1).

When are our experimental findings better than a guess?

Clintin P. Davis-Stober[1] & Jason Dana[2]
[1]*University of Missouri,* [2]*University of Pennsylvania*

We address the issue of estimation accuracy and how it relates to the crisis of confidence in psychology and related areas. By focusing on estimation accuracy, we are attacking the problem at a basic level. We literally ask whether sample means, under typical sample and effect sizes in our discipline, are accurate and reliable enough representations of the true means to be used as evidence for testing psychological theories.

How accurate should our estimates be? Many areas of behavioral research are not sufficiently quantified to precisely answer this question. We approach this question through the use of a benchmark. Consider a fundamental experimental design in which subjects are assigned to different treatment groups, whose means on some dependent measure are then compared to determine whether the experimental treatments had an effect. We present an alternative to sample means that produces random guesses about the direction and magnitude of treatment effects called a random conclusions estimator. Most researchers would agree that using a random conclusions estimator is absurd science. Yet, we show that many areas of behavioral research typically operate at sample and effect sizes for which sample means are less accurate, on average, than our random conclusions estimator. To put a finer point to our argument, it is possible to have a legitimate, un-hacked, statistically significant difference in means while those means reflect the truth worse than our guessing benchmark. Under these conditions, sample means are unreliable across replicates and we shouldn't, a priori, expect findings based on them to replicate.

In this sense, we argue that sufficient estimation accuracy is a pre-potent problem: Before we even begin to place statistical assumptions on data and run significance tests, the quality of our inferences is fundamentally limited by how accurate our estimates are. We discuss what can be done in light of the problem, including setting minimum sample size and/or effect size recommendations based on clearing the guessing benchmark.

How replicable are process measures in decision making? The impact of subtle search costs

Michael Schulte-Mecklenbeck[1], Thorsten Pachur[1], Ryan O Murphy[2], & Ralph Hertwig[1]
[1]*Max Planck Institute for Human Development,* [2]*ETH Zürich*

"One replication is worth a thousand t-tests."

In the field of judgment and decision making in particular, and in psychology in general, replication has been an important means to evaluate the generalizability of research results. We examine how subtle factors in an experimental design can have substantial effects on the replicability of findings concerning the cognitive processes underlying decision making.

One commonly used methodology to examine the cognitive processes in decision making is the process-tracing tool Mouselab, which records movements of a computer mouse on stimuli presented on a computer screen (usually in the form of an information matrix). We examine the replicability of insights using Mouselab in two respects. First, how replicable are individual differences in acquisition patterns? Second, to what extent is the replicability of acquisition patterns affected by subtle factors in the experimental set-up? Specifically, Mouselab offers the researcher the use of two different modes of information acquisition: an information box is opened either by clicking on it (click method) or by moving the mouse over it (mouseover method). Does this seemingly minute difference (either a click is required or not) in method impact the replicability of how people search for information? Does it affect the replicability of people's decide behavior?

In the current study we address these questions in the context of a risky choice task, in which participants were asked to indicate their preferences among two-outcome lottery problems. For each risky choice, participants evaluated the lotteries in terms of their possible outcomes and probabilities. This information was hidden behind boxes but could be revealed either by clicking on the box (group 1: click condition) or by moving the mouse over the box (group 2: mouseover condition). In all other regards, the setup in the two conditions was identical. After a period of three weeks, participants came back to lab and were presented with the identical lottery problems (in a different order).

It emerged that individual differences in acquisition patterns, as captured with Mouselab, are relatively stable between the two sessions. Mouselab thus captures replicable aspects of information processing. Further, although the information presentation in the two acquisition conditions differed only in terms of the subtle costs imposed by having to click (or not) to reveal a piece of information, we find substantial differences between the mouseover and the click conditions. First, participants in the click condition acquired less information, consistent with the idea that clicking incurs higher costs than the mouseover method. Second, we find differences in participants' search direction: in the click condition information search was less option-wise than in the mouseover condition. Finally, using computational modeling we also find differences between the conditions in people's choices. Based on cumulative prospect theory (Tversky & Kahneman, 1992), we find that participants in the click condition were more loss averse, showed lower probability sensitivity, and a lower choice consistency. We contrast the size of the effect of acquisition mode also to the effects due to whether information for an option is presented in a vertical or a horizontal format.

Our results illustrate how relatively minute and so far neglected methodological aspects in process tracing studies can impact the replicability of findings regarding people's decisions and the decision process. We will discuss implications of our results on a methodological level in light of replication over extended periods of time as well as processing costs and default effects when running experiments using the Mouselab tool.

Advancing replicability and theory through replication recipes and replication packages

Mark J. Brandt[1], Hans IJzerman[1], Roger Giner-Sorolla[2], Frank Farach[3], James Grange[4], Jason Geller[5], Jeffrey Spies[6], Anna Van t' Veer[1], & Marco Perugini[7]
[1]Tilburg University, [2]University of Kent, [3]University of Washington, [4]Keele University, [5]Iowa State University, [6]University of Virginia, [7]University of Milano-Bicocca

This talk outlines guidelines for replication attempts, and how researchers should report their study so as to allow future researchers to replicate their work. We start with the premise that direct replications are often extremely difficult in psychology because of differences in participants and contexts. However, just as there is value in replications examining generalizability, there is value in close replications in order to develop a solid and cumulative body of evidence. We have developed a replication "recipe" to facilitate close and convincing replication attempts. The replication recipe aims to standardize the criteria for a convincing replications including as faithfully as possible recreating the original study (and keeping track of differences). This includes using high-powered studies, checking the study's assumptions in new contexts, pre-registering the study, and methods for evaluating and reporting the replication. By identifying the different facets (sample, culture, lab context, etc.) on which the replication may differ, it allows researchers to identify whether their replication is "close" or "conceptual". Our replication recipe can be used by established researchers, teachers, and students to conduct meaningful replication studies and integrate replications into their scholarly habits. We also suggest that replication packages--collections of the data, materials, code, and other materials necessary for another researcher to replicate the analyses and procedures of a study--will serve to make the replication recipe easier to follow. Sharing materials in this manner, as opposed to the current standard of summarizing the methods ins a manuscript, will make it is easier to closely recreate the original study. In summary, via the replication recipe and packages, we aim to make replications easier, more convincing, and more likely to advance theory.

Data-blind peer review

Christoph Stahl
University of Cologne

The current discussion about methodological controversies indicates some problems not only with current research practices, but also with incentives. For any individual researcher, the successful publication of his or her study is arguably the central and most important incentive. In current publication decisions, the results of a study figure importantly. The results of empirical studies therefore have two roles: Not only are they fundamental to help answer our research questions and to further our knowledge; simultaneously, they are the means by which we advance our individual careers. The latter fact invites researchers to use the available (questionable but legitimate) research practices to their advantage (e.g., researchers' degrees of freedom, hypothesizing after the results are known). As a consequence, our empirical findings – the core of our science – are under pressure because they serve two (potentially conflicting) goals: the growth of knowledge, as well as the scientist's career, via their influence on publication decisions.

In my contribution, I will therefore argue for a data-blind peer review process (Greve, Bröder, & Erdfelder, in press) that removes the influence of results on publication decisions. As a first step toward this goal, I propose a new type of publication, the registered report (Chambers, 2013). In this model, similar to grant proposals, manuscripts will be reviewed (and accepted for publication) before data collection. Submissions would include sections on introduction, methods, analysis protocol, and perhaps pilot data, on a level of detail that would allow for independent replication of data collection and analysis. Decisions would be based on theoretical relevance and on methodological adequacy and rigor (not on the results or the cohesiveness of the narrative). In this model, the results would no longer be in the focus of editorial decisions: Studies would be published that promise to be interesting no matter how the results turn out, and researchers would instead be rewarded for their theoretical and methodological contributions.

The revised reward structure would reduce the incentives to use questionable research practices, and thereby render many measures to control their use obsolete. Importantly, registered reports would be well-suited for high-powered replication studies on effects still under debate, as well as for the publication of null effects. Additional advantages as well as some limitations will be discussed. In sum, if empirical findings could no longer affect publication decisions, they would be allowed to more exclusively serve the growth of knowledge.

Psychologists are open to change, yet wary of rules

Mirjam Jenny[1], Heather Fuchs[2], & Susann Fiedler[3]
[1]Max Planck Institute for Human Development, [2]University of Cologne, & [3]Max Planck Institute for Research on Collective Goods

Recent scandals in psychology as well as new discussions of old methodological problems in certain psychological disciplines have made it clear that psychologists must change the way they conduct and report their research. This presentation gives an overview over the current debate in the psychological field before focussing on the publishing process. One article recently published in Psychological Science proposing six requirements for researchers concerning data collection and reporting practices as well as four guidelines for reviewers aimed at improving the publication process has recently received much attention (Simmons, Nelson, & Simonsohn, 2011). We surveyed 1,292 psychologists to address two questions: Do psychologists support these concrete changes to data collection, reporting, and publication practices, and if not, what are their reasons? Respondents also indicated the percentage of print and online journal space that should be dedicated to novel studies and direct replications as well as the percentage of published psychological research that they believed would be confirmed if direct replications were conducted. We found that psychologists are generally open to change. Five requirements for researchers and three guidelines for reviewers were supported as standards of good practice, whereas one requirement was even supported as a publication condition. Psychologists appear to be less in favor of mandatory conditions of publication than standards of good practice.We conclude that the proposal made by Simmons, Nelson & Simonsohn (2011) is a starting point for such standards.

The empirical content of theories in judgment and decision making: Shortcomings and remedies

Andreas Glöckner & Tilmann Betsch
University of Göttingen

According to Karl Popper, we can tell good theories from poor ones by assessing their empirical content (empirischer Gehalt), which basically reflects how much information they convey concerning the world. “The empirical content of a statement increases with its degree of falsifiability: the more a statement forbids, the more it says about the world of experience.” Two criteria to evaluate the empirical content of a theory are their level of universality (Allgemeinheit) and their degree of precision (Bestimmtheit). The former specifies how many situations it can be applied to. The latter refers to the specificity in prediction, that is, how many subclasses of realizations it allows. We conduct an analysis of the empirical content of theories in Judgment and Decision Making (JDM) and identify the challenges in theory formulation for different classes of models. Elaborating on classic Popperian ideas, we suggest some guidelines for publication of theoretical work.

How Important is Statistical Hypothesis Testing for the Quality of Science?

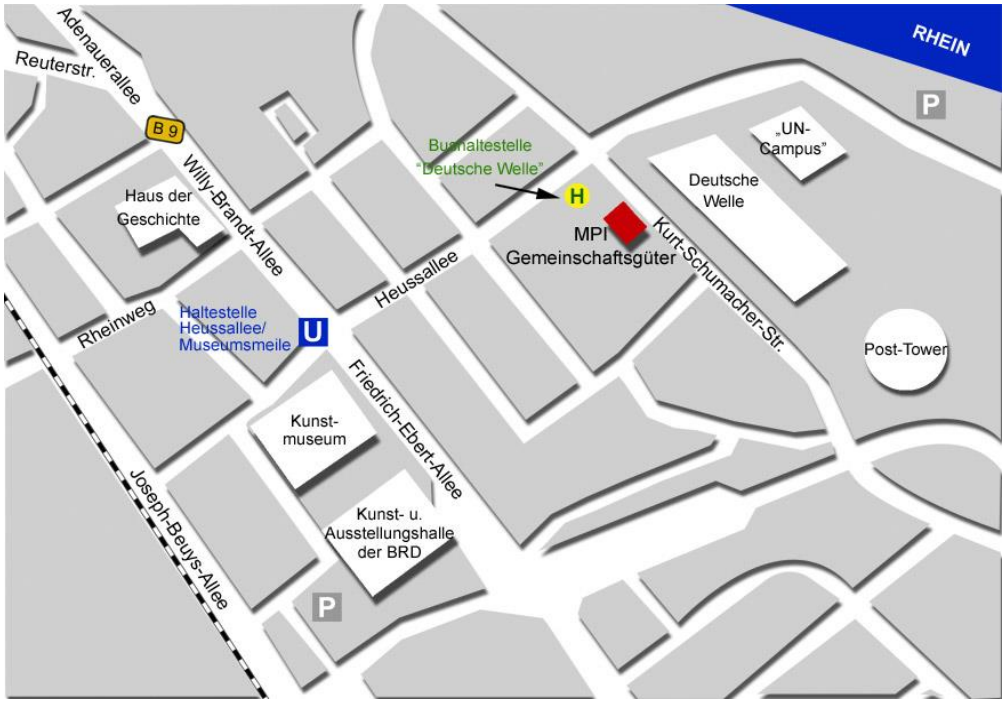
Klaus Fiedler
University of Heidelberg

In the current debate about replicability and usability of scientific findings and about ways of improving the quality of behavioral research, a key role is commonly attributed to the rigor of statistical methods. Although I am myself a fan of the creative use of statistical tools, I doubt that statistical analysis can contribute much to progress in science. The history of behavioral science in general and of research in judgment and decision making in particular does not reveal any strong examples for beneficial effects of stricter hypothesis testing on the quality of science.

To be sure, there are many examples of statistical models (e.g., signal detection) that have led to great innovations. However, strict limits imposed on alpha errors or effect sizes in data analysis cannot be expected to produce ground-breaking new insights. Whoever has participated in a logic course, or is familiar with the Wason card-problem, knows that a test of the hypothesis “if p, then q” does not tell us anything on the (conditional or unconditional) likelihood of p. By analogy, whether a statistical test of a hypothesis q suggested by theory p is significant or not – any outcome may be due to other influences on q (not covered by p) or to countless boundary conditions introduced by the operationalization of q. Because there are always multiple causal factors besides p that also affect q, even the most reliable evidence on q does not enable a reverse inference to p (rather than p’, p’’ etc.) as the ultimate cause. Developing a comprehensive theoretical framework within which competing theories can be evaluated is therefore more important than merely improving the habits of statistical testing. Also, no cost-benefit analysis has ever shown that false positives are more expensive or more irresponsible than false negatives. To illustrate the over-estimation of statistics in behavioral science, I will provide examples of unwarranted inferences from Bayesian updating, mediation analysis, and paramorphic modeling.

Location

How to find us
The Institute's address is Kurt-Schumacher-Str. 10, 53113 Bonn



By taxi:
Taxi fares are about € 12 from the city center and about € 35 from Cologne/Bonn Airport. The central phone number for Bonn taxis is 555 555.

By public transport:
From the central station, take bus number 610 towards Bad Godesberg-Rheinallee. Get off at Bonn-Gronau-Deutsche Welle. This trip takes about 10 minutes. The institute is opposite the Deutsche Welle building.

An alternative is to take the streetcar (U-Bahn). Take U-Bahn 66 towards Bonn Ramersdorf or U-Bahn 16, 63 or 67 towards Bad Godesberg. Get off at Heussallee/Museumsmeile. Walk along Heussallee following the signs for the "Deutsche Welle". Turn right into Kurt-Schumacher-Str. The institute is on the right-hand side of the street, approximately 50 meters from the corner. This trip takes around 20 minutes.

From Cologne/Bonn Airport, first take bus number 670 to the central train station, then bus number 610 as described above. (Airport buses leave approximately every 20 minutes. No advance booking is necessary).

By car:

From Frankfurt:

From the A 59, take the A 562 at Kreuz Bonn-Ost. Follow it over the Konrad Adenauer Bridge. At the first exit after the bridge turn right onto Franz-Joseph-Strauss-Allee. Take the first left (Sträßchensweg), and follow the street straight ahead. This street becomes Kurt-Schumacher-Str. The institute will be on your left, opposite the Deutsche Welle.

From Cologne:

From the A 555, take the A 565 towards Koblenz. Exit at "BN-Poppelsdorf/Bad-Godesberg". Follow the Reuterstraße onto Willy-Brandt-Allee turning left into Heussallee and then taking the second street on your right, the Kurt-Schumacher-Str.