Ludwig-Maximilians-Universität München

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)
am Munich Center of the Learning Sciences
der Ludwig-Maximilians-Universität München

How do researchers evaluate statistical evidence when drawing inferences from data?

Arianne Herrera-Bennett (*PhD candidate*)

Munich. 03. 06. 2019
Date of oral defense (27.09.2019)

Supervisory Team:
Prof. Dr. Moritz Heene (*1st supervisor*)
Prof. Dr. Stefan Ufer (*2nd supervisor*)
Dr. Daniël Lakens (*3rd supervisor*)

Acknowledgments

I would like to sincerely thank my supervisory team, Prof. Dr. Moritz Heene, Prof. Dr. Stefan Ufer, and Dr. Daniël Lakens, for all their incredible support and guidance throughout the last three years, and with whom it has been a joy to learn from and collaborate with.

The research presented in this work was supported by the Elite Network of Bavaria [Project number: K-GS-2012-209]. I would like to extend my gratitude to the opportunities (conference attendance, incubator stay, international knowledge exchange) made possible by the ENB.

Abstract

(Dissertation)

The current dissertation asks two core questions: *How* do researchers evaluate statistical evidence when drawing inferences from data? And how can we *improve* this process of statistical inference-making among researchers in the field of psychology? In order to gain some insights and possible answers to these questions, two projects were carried out. The first (**Project 1**) took a more literal approach and assessed the effectiveness of an 8-week massive open online course (MOOC) at reducing individuals' propensity to fall prey to common statistical misconceptions that have been observed as widespread amid the research community, and thus assumed to be resistant to change. Not only did findings indicate that misconceptions (concerning *p*-values, confidence intervals, and Bayes factors), were able to be improved in terms of immediate learning, it was also found that this learning could be maintained across the 8-week timeframe (**Study 1**). Moreover, additional instructional support which served to explicitly clarify and train individuals to recognize *p*-value misconceptions (**Study 2**) further bolstered this improvement. The second project (**Project 2**) took a meta-scientific approach to these questions, exploring how the concepts of model generalizability and replicability might relate in practice: Specifically, by re-analyzing the replication data sets of the recent 2018 replication project (Camerer et al., 2018), using a repeated *k*-fold cross validation technique, we could test whether model prediction accuracy of a study could positively predict whether it successfully replicated or not. Preliminary results were in line with our intuitions, and corroborated the notion that strength of initial evidence – which may be captured in multiple ways – may be an important determinant of study replicability. Taken together, the two projects speak to the merits of fostering a more thorough and nuanced understanding of statistical data among researchers, in order to move away from the traditional rote way of teaching and using statistics, and toward a more comprehensive and thoughtful approach to drawing well-founded and meaningful statistical inferences from data.

Word count (dissertation, references incl.) = approx. 46,784

Table of contents

INTRODUCTION

1. General introduction

The ability to draw meaningful and accurate statistical inferences from data is a key component in how researchers make sense of research findings, interpret their meaning, and in turn offer implications and conclusions. The trouble is that within disciplines, such as psychology, inferential statistics, which are inductive by nature (i.e. involving the estimation of a population parameter from a sample-based estimator), are necessarily based on probabilities, and thus always entail some degree of uncertainty. Therefore, unlike deduction that results in *valid* versus *invalid* arguments (for details, see **section 2.1**), induction relies on corroborating evidence to *strengthen* versus *weaken* claims. Beyond the fact that dealing with uncertainty adds complexity when evaluating statistical evidence, a central stumbling block when it comes to psychological research in particular – the focus of the current dissertation – is the inordinate use of null hypothesis significance testing (NHST), a framework for 'statistical inference-making' (i.e. drawing statistical inferences from data) which seems to impart the illusion that one can accept or reject hypotheses or theories with certainty. Paired with the long-standing criticism that NHST, i.e. use of *p*-values, promotes rote mechanical and dichotomized thinking (for review of NHST controversy, see Nickerson, 2000), a consequence that follows from the overreliance on and misuse of significance testing, are widespread inconsistencies between proper statistical theory and faulty research practice. As a result, critics of the NHST approach have pushed for statistical reforms advocating the use of alternative techniques (a theme which will be carried throughout the dissertation), such as methods which encourage estimation thinking, e.g.,

confidence intervals (CIs), effect sizes, meta-analysis, or use of Bayesian statistics (e.g., Bakan, 1966; Falk & Greenbaum, 1995; Cohen, 1990; Rozeboom, 1997; Wagenmakers, 2007; Dienes, 2011; Morey, Romeijn, & Rouder, 2016).

In addition to issues stemming from misunderstandings or misuses of the NHST approach itself, the competitive academic culture, within which psychology research is conducted, is also commonly blamed for incentivizing researchers to inflate claims, fall prey to cognitive biases, or appeal to research practices that are questionable and self-serving. Consequently, when it comes to drawing statistical inferences, over-generalizations and inflated claims about the nature of observed effects can perpetuate false trust in the existence of phenomena, which is particularly conflicting given contradictory findings (e.g., negative or null results, or failures to replicate). For these reasons, it is not surprising that contemporary science finds itself in a 'replication crisis', and that calls for system-wide changes in research practices and improved methods are ever-growing (e.g., Wasserstein & Lazar, 2016; Amrhein, Greenland, & McShane, 2019).

As a response to these calls, the current dissertation proposes to investigate the following questions: *How* do researchers evaluate statistical evidence when drawing inferences from data? And how can we *improve* the process of statistical inference-making in order to foster accurate and meaningful inferences from data? While improving the practice of inferential statistics among researchers is only one component in fostering more meaningful outcomes from scientific investigations, it is nonetheless one of the most heated points of contention within the scientific community on account of the very fact that the culture of research puts so much precedence on hypothesis-supporting (i.e. positive) study results. In fact, the misuse of statistics often stems from a general (or intentional) ignorance about the complexities that surround the process of engaging in valid statistical reasoning, amounting to more significance-chasing at the expense of

critical thinking. Fostering effective statistical thinking in researchers is thus a multi-faceted undertaking, requiring (but not limited to): the skill to move beyond rote use of NHST and ask questions that are relevant to one's given research question(s) and goals, recognizing and identifying inherent layers of uncertainty that exist within investigations, replacing dichotomous thinking with a more nuanced appraisal of data, and overcoming biases toward significant results. While perhaps unrealistically optimistic, shifting away from the routine use of significance thresholds, and adopting a more comprehensive understanding of the strengths and limitations of a variety of methods, should in theory provide a better foundation for thoughtful research. Though broad, this underlying assumption plays a role in each of the dissertation projects presented, of which the details will be elucidated below (see **section 1.2**).

Given the complex nature surrounding the questions posed above, it makes sense to briefly digress and explicitly appeal to the idea of multiple levels of research as well as inference when considering how individual researchers evaluate data (see **Figure 1** below). In other words, when it comes to understanding how individuals engage in the process of statistical inference-making, sources of observed differences can originate at the level of the individual researchers themselves, such as individual-level differences in grasping certain statistical concepts. On the other hand, as described above, many sources of influence exist as a result of more top-down influences, i.e. at the level of the research community, such as instructional norms or incentive structures (see **Figure 1A**). Thus, research like ours which seeks to investigate how individuals draw inferences, should also consider accounting for the broader context within which statistical inferences are drawn. In a similar vein, statistical inferences do not reside on a single level of analysis (see **Figure 1B**): While the psychological research community has put an "undue emphasis on significance levels" (Ioannidis & Trikalinos, 2007, p. 245), and thus research on *p*-

value interpretations is warranted, interpreting the outcome of a specific statistical test (e.g., *p*-value) is only one component of how researchers are likely to interpret the outcome of an entire study, let alone the meaning of that one study amid the full body of literature within which it belongs. Therefore, once again, arriving at a more comprehensive answer to the overarching dissertation questions posed should necessarily consider these different levels of analysis.

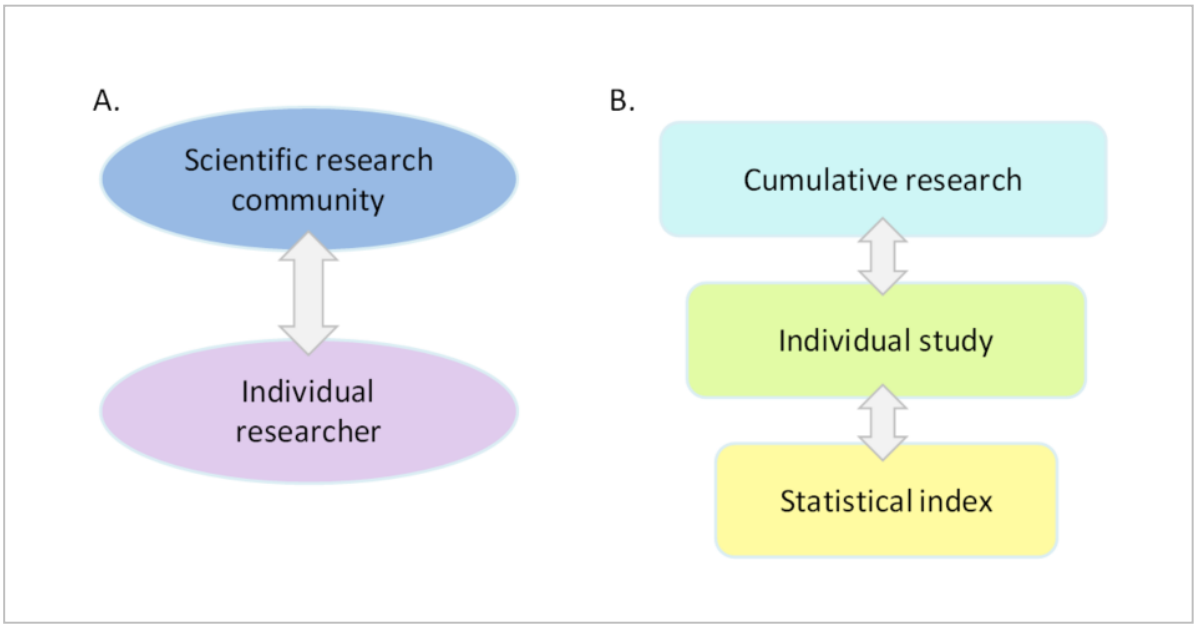

Figure 1. A. Levels of research: Individual researchers vs. the broader scientific research community. **B. Levels of inference:** Drawing statistical inferences can occur at the different levels: level of individual statistical indices, in terms of the outcome of individual studies, or at the level of the full body of research findings taken together (cumulative research).

1.1. Structured overview

The current dissertation presents two projects, both broadly geared at better understanding *how* researchers draw inferences from statistical data, within the area of psychological research[1]. Moreover, in keeping with the continued push toward statistical reforms within psychology, central to both thesis projects is also the goal of developing and assessing means to *improve* the practice of drawing statistical inferences. Specifically, each of the projects bears relevance to one of two core areas of contention that have arisen in statistical reforms, namely the NHST controversy and the 'replication crisis'. The first project (see **section 1.1.1** for *brief introduction*) challenges the idea that certain statistical misconceptions, primarily widespread p-value fallacies, are fixed or resistant to change, and empirically tests whether they can be systematically improved. The second project (see **section 1.1.2** for *brief introduction*) is centered around investigating the nature of replicability, and tests how the assumption of cross-validation techniques, as an index of generalizability, holds up in practice when used to predict the replicability of real data sets. To help give more context, both projects are very briefly described in some more words below, after which a structured overview of the dissertation's core line of argumentation (see **section 1.1.3**) is outlined.

1.1.1 Project 1 brief introduction. The first project investigated statistical inference-making at the level of individual course learners enrolled in an 8-week massive open online course (MOOC) on *"Improving your statistical inferences"*. Specifically, in **Study 1**, baseline misconception rates, as well as improvement rates, for the concepts of *p*-values, confidence intervals (CIs), and Bayes factors (BFs), are assessed across the 8-week course. Building directly

[1] **Note:** While the applicability of concepts presented are not being argued as isolated to the discipline of psychology, for the purpose of the dissertation, the framework and discussions are limited to this context.

off of past research, items target common misconceptions that have been observed as particularly prevalent among psychological researchers. Additionally, as a novel contribution, baseline accuracy rates for *p*-value misconceptions are compared under two separate contexts, namely: whether *p*-value statements refer to a significant outcome (i.e. $p = .001$) or not ($p = .30$). In this way, though only indirectly, we attempted to address whether certain top-down biases (at the level of research incentives) contribute to individuals' propensity to fall prey to certain misconceptions.

In **Study 2**, we assessed how additional instructional support and training, directed at explicitly pinpointing common *p*-value misconceptions, could potentially further improve rates of learning. **Note:** Regarding the use of the term 'misconception', the statistical misconceptions literature has not so strictly delineated its definition from other related terms, thus for the purposes of the present dissertation, the term 'misconception' refers "to any sort of fallacies, misunderstandings, misuses, or misinterpretations of [statistical] concepts, provided that they result in a documented systematic pattern of error" (for review, see Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007, p. 99). In other words, consistent with the literature, these terms are used interchangeably; one core reason being that the current work is not making claims about the route source of the misconceptions being measured, but rather leaving open the possibility that they originate e.g., from innate misguided intuitions or merely a lack of knowledge (see **section 4.1** for details).

1.1.2. Project 2 brief introduction. The second project investigated how researchers make sense of data primarily at the level of individual study outcomes. Specifically, it consisted of a large-scale re-analysis of the most recent Social Sciences Replication Project (SSRP) by Camerer and colleagues (2018), applying a cross-validation (CV) technique (i.e. five-repeated

10-fold CV) in order to explore whether measures of model generalizability can be used as predictors of replicability. In other words: Can certain individual study features (i.e. CV indices) be used to predict future replication success. In this sense, the project operates at all three levels of analysis (**Figure 1B**), bridging features at the level of specific statistical indices with prospective assumptions at the level of cumulative research. While presumably abstract, making this leap from statistical index to a generalized claim about replicability is not unheard of. In fact, falsely assuming that the *p*-value significance of a study is directly predictive of replication success is among one of the most common *p*-value fallacies observed (see *replication* fallacy, **Box 1**, p. 23). Therefore, improving our understanding of replicability, not only as it concerns comparing original and replication outcomes, but also in terms of statistical predictors, may provide a fuller framework through which to understand the limits of how much can be inferred from a single study finding within the context of cumulative research. **Note:** Here, 'replicability' refers to observing the same finding in a new set of data (replication data set) when using either a close or dissimilar methodological approach as the original study (i.e. direct vs. conceptual replications); 'reproducibility' refers to arriving at the same results by running the same analysis on the original data set. This dissertation is primarily concerned with replicability.

Though quite disparate, these two projects do share some commonalities: Beyond the fact that they both touch upon the topic of replication, there is more broadly shared theme is the assumption that capitalizing upon more pieces of statistical information, rather than relying on one kind of evidence or one routine approach, should provide a better foundation upon which to draw thoughtful conclusions from observed data. Of course, it should not simply be argued that more is always better; however (as will be discussed in the coming sections), there are fairly compelling grounds to believe that the practice of inferential statistics, when it comes to

reporting on psychological research, is currently greatly limited in its application, and in turn has left researchers ill-equipped to move beyond a restrictive style of categorical thinking. Therefore, beyond walking through the state of research concerning topics related to inferential statistics (e.g., the misuse of statistics, statistical reforms, etc.), some sections will also pull in some theoretical literature that may provide a broader framework through which to conceptualize how to facilitate a shift in thinking, i.e. away from dichotomous framing, and toward a more nuanced process of drawing statistical inferences.

1.1.3. Dissertation overview: Line of argumentation. The present dissertation is structured as follows: The rest of the introduction is organized into three overarching sections (**sections 2, 3, & 4)** outlined below, followed by the inserted full manuscripts (**Projects 1 & 2)** introduced above, after which a final **General Discussion** is offered (**sections 5 & 6**). The following three pages outline the overarching line of argumentation carried throughout the thesis.

First, the section **'Statistical thinking: Reasoning under uncertainty'** (**section 2**) establishes the concept of inductive inferences and the role of uncertainty (**section 2.1**) commonly overlooked when individuals are faced with probabilistic statements. This concept is important to address when assessing how researchers draw inferences in psychological research, especially given the overreliance on NHST, as the process of using probabilities to estimate a population parameter from a sample-based estimator constitutes the act of making an inductive generalization (**section 2.2**). Thus, should instruction and/or use of significance testing commonly fail to communicate this inherent level of uncertainty, and instead promote a rote approach to statistical inferences that results in artificial and misleading categorizations of effects (e.g., significant or not; true or false), then resulting claims should hardly be indicative of rigorous or well-founded statistical understanding. As such, the idea of effective statistical

thinking is discussed briefly (**section 2.3**) as constituting a skill that goes beyond the ability to apply statistical techniques, but also in critically asking relevant and more nuanced questions that account for research uncertainty.

The next section **'Contemporary science, controversies, & statistical reforms'** (**section 3**) tackles two core controversies that have had predominant roles in contemporary science and statistical reforms, and which each of the dissertation projects bear specific relevance on, namely: The misuse of significance testing and the prevalence of *p*-value misconceptions (**section 3.1**) relates to **Project 1**, and researchers' response to the replication crisis (**section 3.2**), relates to **Project 2**. Interestingly, as already alluded to above, it is possible to speculate on how misunderstandings at the level of statistical indices (e.g., *p*-value fallacies, CI misconceptions) in fact carry-over when researchers judge cumulative evidence: Exemplified perhaps best by the *replication* fallacy or the widely held confidence-level misconception (see **section 3.2**), lack of proper understanding of these frequentist concepts (i.e. *p*-values, CIs) is likely to have contributed to how researchers perceive and interpret failed replications, and possibly why some were inclined to discredit the findings of the initial large-scale replication project (Open Science Collaboration, 2015) as evidence of a crisis in psychology. In light of such speculations, a case is made for the need for more meta-research, i.e. the empirical study of research itself (**section 3.3**), when addressing questions that operate at multiple levels of analysis, as in the case of investigating how researchers statistically reason (e.g., how they make sense of the concept of replicability). Finally, accounting for the culture of research within which researchers routinely draw inferences (**section 3.4**) is similarly important in order to more comprehensively gain an appreciation of which sources of more top-down biases (e.g., publish-or-perish culture) may overshadow correct use – or exacerbate misuse – of inferential statistics.

The last section **'Fostering statistical & meta-scientific thinking'** (**section 4**) shifts focus away from sources of statistical misunderstandings, and toward the idea of how improvements in statistical inference-making might be fostered, including the notion of bettering one's ability to formulate salient questions when approaching research, and optimally assessing and weighting multiple (rather than single) pieces of evidence. In order to better conceptualize what it might mean to foster improvements in individuals' skills and understanding, two frameworks are argued to offer some insights. First, I appeal briefly to the conceptual change literature (**section 4.1**) and the notion of the novice-expert dichotomy with regard to the acquisition of expertise. According to this tradition, the prevalence of *p*-value misconceptions, among both students and experts (e.g., statisticians or methodology instructors), would typically suggest that these misconceptions continue to exist because they withstand proper instruction, possibly reinforced by past and/or common real-world experiences. The limitation in applying this framework to the understanding of NHST misuse (as underscored in **sections 2.1** and **3.1**) is that significance testing has routinely been taught in a formulaic fashion, i.e. as a rote ritual that encourages limited dichotomous thinking. Moreover, very little research has investigated interventions to improve *p*-value misconceptions. Therefore, norms of instruction and routine misuse of NHST, in combination with the current state of research, make it tough to judge whether the widespread misuse of frequentist statistics speaks to some innate problem in understanding its correct use, or simply to a lack of knowledge or effective instruction. Put differently, in assessing whether *p*-value fallacies (or other statistical misconceptions) can be improved (i.e. **Project 1**), determining whether these improvements are a product of overcoming naïve conceptions versus just providing the correct tools to infer the correct meaning – or realistically some combination of both – is a tricky task to reconcile.

Bearing this in mind, I next appeal to the problem-solving literature (**section 4.2**) particularly as it pertains to the concepts of effective problem formulation, fixation, and selecting the appropriate tool or strategy to solve a problem. According to Gestalt tradition, *fixation*, i.e. the tendency to approach a problem in a particular or similar way, typically as a result of previous familiarity or experience, needs to be overcome by fundamentally restructuring one's problem space, sometimes referred to as *breaking frame*. Regarding the extent to which NHST has been institutionalized among psychology researchers, I would argue that this need to fundamentally restructure the problem space can be applied more holistically to the current practice of drawing inferences in psychology: In sum, reconfiguring the problem constraints that characterize the process of drawing inferences – so that researchers' goal states, and tools to draw inferences, are not fixated on significant results – may be one way to more globally conceptualize the presence of multiple sources of influence simultaneously contributing to how or why certain misconceptions are so persistent. Though this use of the Gestalt framework is more conceptual in nature, I would still contend that (when taken alongside the conceptual change literature), it can offer complimentary implications for instruction, that is: While one emphasizes the use of formal instruction to refine one's understanding of statistical concepts, the other stresses the need to also act concurrently on other sources of influence (e.g., cognitive biases, incentive structures) that may otherwise override the success of proper instruction.

After presenting both manuscripts (introduced above, see **sections 1.1.1 & 1.1.2**) the dissertation finishes with a general discussion that brings together some concepts common to both projects, and speculates on additional implications for instruction, as well as possible avenues of research, in order to further the understanding and improvement of drawing statistical inferences from data, among researchers within psychology.

2. STATISTICAL THINKING: REASONING UNDER UNCERTAINTY

2.1. Statistics: The science of uncertain inference

The practice of inferential statistics in scientific research, namely as it concerns the discipline of psychology, has been heavily grounded – if not fundamentally defined – by the frequentist tradition of defining probabilities. Such an inference model has its foundations in "inductive philosophy" (Mayo & Cox, 2006, p. 77) and presupposes a "logic of inductive inference" (Fisher, 1935, p.39; Neyman, 1955). In other words, unlike deductive reasoning, which constitutes analytic derivation (i.e. the premises of the argument already, in the logical sense, contain or entail the conclusion; Greenland, 1998), induction attempts to reason from particulars to generals, and as such has been characterized as a logic of evidential support whereby the *strength* (rather than the *validity*) of an argument is a function of the amount of corroborating evidence. In this sense, these two modes of human reasoning (as distinguished by logicians), differ critically in the extent to which conclusions can be drawn with certainty: Only in the case of a valid deductive argument (e.g., *modus tollens*) do the premises necessarily logically guarantee its conclusion. By contrast, conclusions that follow from inductive inferences cannot be guaranteed to be true.

Specifically in the case of null hypothesis significance testing (NHST), one major source of confusion, which has led researchers to believe that inductive inferences can also be drawn with certainty, is what Falk and Greenbaum (1995) deemed "the illusion of probabilistic proof by contradiction" (p. 76). In essence, this results from a basic probabilistic misinterpretation:

inappropriately applying the use of the (otherwise valid) *modus tollens* conditional argument form (i.e. *if p then q, not-q, therefore not-p*) when interpreting probabilistic statements. In other words, while *modus tollens* follows the structure of a *'reductio ad absurdum* proof' (i.e. proof by contradiction), whereby the rejection of the antecedent (*p*) follows from the contradiction of the consequent (*q*), this argument form is only valid given categorical statements, i.e. in this case, where the outcome contradicted (*"not-q"*) is absolute, not just highly improbable (i.e. $p < .05$). Herein lies the critical difference: When this logical argument form is inappropriately applied to the NHST context, then one might be tempted to assume that observing a highly improbable outcome (let $p = .001$) is analogous to arriving at an outcome so 'absurd' given the original premise or assumption (here, H_0) such that it can justify the rejection of that (null) assumption. In fact, because a probability, no matter how small, is never zero, "the probabilistic counterpart of that logical deduction does not hold" (Falk & Greenbaum, 1995, p. 78).

Even though it is common practice for researchers to appeal thresholds of improbability (i.e. p = alpha) as a decision criteria to reject or abandon hypotheses, it remains that statistics has been characterized as "the science of uncertain inference" (Lehmann, 1993, p. 202): When it comes to probability theory and inferential statistics, "such inferences we recognize to be *uncertain* inferences" (Fisher, 1935, p. 39).

2.2. Statistical inferences: From sample to population

"Probable evidence, in its very nature, affords but an imperfect kind of information."
(Butler, 1736)

In psychological research, inductive inferences take the form of statistical generalizations, using sample-based estimators to draw inferences about population parameters.

Consequently, "as in all inductive inferences, we cannot establish that the statistical generalization is true with absolute certainty" (Innabi & Jordan, 1999, p. 188). This is because, beyond probabilistic misinterpretations (like the one described above; **section 2.1**), experimental data that pertain to probabilistic data-generating models, as in the case of psychology, are necessarily stochastic, i.e. possessing some inherent randomness. As such, a critical issue that accompanies the process of inferring generalized claims from observational data is that a sample, by definition, contains incomplete and/or potentially biased (i.e. non-random) information. Understanding how the relationship between sample and population varies as a function of sample properties (e.g., size, variability, representativeness) thus constitutes one core component leading to meaningful generalizations (see e.g., Sotos et al., 2007), and why the ability to draw inferences from data has been considered as an aspect of critical thinking (Ennis, 1985; Innabi & Jordan, 1999).

Although such a concept should be neither controversial, nor particularly tricky to grasp, neglecting to account for sampling variability (i.e. sheer random variation) within data, let alone potential sources of systematic biases across data (e.g., inflated effect sizes), might partly explain why researchers hold falsely optimistic impressions of the expected consistency or generalizability of effects. While this point will be further elaborated in a later section (see **section 3.2**), this idea relates to both projects of the current dissertation. Where in the first project, statistical inferences are being made at the level of specific statistical indices (e.g., *p*-values, CIs), misconceptions commonly arise when trying to infer beyond the specific sample data in order to judge the outcome of future studies. Specifically, in the case of *p*-values, individuals commonly fall prey to the *replication* fallacy, the false belief that significant studies necessarily have a high probability of replicating (see **Box 1**, p. 23); in the case of CIs,

individuals commonly confuse a 95% CI with the erroneous belief that it will "on average capture 95% of replication means" (Cumming, Williams, & Fidler, 2004, p. 299). Both instances of misunderstanding appear to assume that a single statistical index, or the outcome of a single study, should be sufficient or representative enough of the population effect to serve as predictive criteria of future studies to come (which we know is not the case).

This idea is very much consistent with the second project, which expressly explores potential predictors of study replicability, but through the use of real replication data. Here, while we initially narrow in on a specific statistical concept (i.e. model generalizability) in order to potentially infer the likelihood that individual studies will replicate, the overarching project goal is to harness multiple pieces of information in combination to hopefully yield a more accurate predictive model of future replication success. With that said, given the nature of real data, e.g., potential effect size inflation or selective publishing of the original studies, even more sources of variation need to be accounted for before trying to make sense of why certain statistical indices may be observed as successful predictors of replicability.

All in all, one important point worth stressing is that both the act of drawing statistical inferences, as well as research geared at better understanding and improving this process, are neither linear undertakings: While drawing statistical inferences might subsume some general mathematical knowledge and competence (such as understanding probability rules; see e.g., Chiesi & Primi, 2010), or the ability to apply statistical techniques, the practice of inferential statistics cannot be reduced to any one given isolated skill or process. Rather, it operates necessarily within a broader framework of inference, in which any given sample-based estimate must be evaluated in the context of the entire research cycle, before generalized conjectures can be offered about the population parameter in question. In a similar vein, assessing how

researchers make sense of statistical concepts, and across different levels of analyses (when applicable), should hinge on pooling from multiple research perspectives (e.g., appealing to relevant aspects of both dissertation projects, for instance, when assessing the concept of replication; more on this last point in **section 5.2.1** of the **General Discussion**).

2.3. Effective statistical thinking

As the last section stressed, drawing inferences from data involves more than performing statistical methods and interpreting statistical indices: "Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as p-values or intervals" (Amrhein et al., 2019, p. 307). The ability to factor in multiple sources of statistical evidence and levels of understanding – including "the omnipresent nature of variation" (Ben-Zvi & Garfield, 2004, p. 7) – has been summarized under the definition of *statistical thinking* as follows (for a comprehensive overview, see Ben-Zvi & Garfield, 2004):

> "Statistical thinking involves an understanding of the nature of sampling, how we make inferences from samples to populations, and why designed experiments are needed in order to establish causation. It includes an understanding of how models are used to simulate random phenomena, how data are produced to estimate probabilities, and how, when, and why existing inferential tools can be used to aid an investigative process. Statistical thinking also includes being able to understand and utilize the context of a problem in forming investigations and drawing conclusions, and recognizing and understanding the entire process (from question posing to data collection to choosing analyses to testing assumptions, etc.)." (Ben-Zvi & Garfield, 2004, p. 7)

For the purposes of the current dissertation, 'effective statistical thinking' borrows directly from this definition and simply involves engaging in this style of statistical thinking in an effective manner so as to most optimally weight various contributing pieces of information with regard to the context of inference. Importantly, this also entails recognizing when the context of information is too limited or simply insufficient to warrant drawing conclusions.

Part of drawing well-founded inferences through effective statistical thinking, involves not only weighting different sources of evidence, and providing answers accordingly (with a specified level of uncertainty), but also considering first and foremost which questions are relevant to ask in the first place: "Researchers often ask multiple different questions at different phases of a research project, and the questions they ask depend on the field, the specific study, previous knowledge, and their philosophy of science" (Lakens, 2019, p. 5). Because the nature of research questions within psychology (and the specifics surrounding their investigation) is diverse, no set of invariant criteria exists to distinguish between which questions should or should not be asked, and as such it is up to the researcher to make sense of and pinpoint those salient to his or her specific research aim(s). In this sense, and in no uncertain terms: "Judgment is part of the art of statistics" (Gigerenzer, 2004, p. 604).

While far from exhaustive, some general questions that a researcher might ask (see Mayo & Cox, 2006; Lakens, 2019) include:

- *"What would falsify my hypothesis?"*
- *"What may be justifiably inferred?"*
- *"How can the gap between available data and theoretical claims be bridged reliably?"*

Or perhaps more specifically in relation to inductive reasoning: *"What is the nature and role of probabilistic concepts, methods, and models in making inferences in the face of limited data, uncertainty and error?"* (Mayo & Cox, 2006, p. 78).

This first overarching section of the dissertation outlined some core complexities of inductive reasoning and inferential statistics (e.g., reasoning and generalizing under uncertainty when using probabilities to move from sample to population), as well as components of effective statistical thinking (e.g., weighing multiple pieces of evidence and fostering the ability to ask nuanced and salient questions). All of these ideas are summarized fairly astutely by this final question provided by Mayo & Cox (2006) directly above. The next section furthers the discussion on the need to foster better inferential skills among psychology researchers, by highlighting two ever-prevalent issues which continue to plague the research community: namely, long-standing criticisms and high misconception rates surrounding the use of NHST, and the replication crisis.

3. CONTEMPORARY SCIENCE, CONTROVERSIES, & STATISTICAL REFORMS

The next section tackles two core controversies that have had predominant roles in contemporary science and statistical reforms within psychology. Each warrants that more empirical research be developed and carried out in order to better understand their respective complexities, and in turn possible avenues for subsequent improvement. The current dissertation projects aim to tackle these goals. First, backgrounds of each of the controversies, as well as how

they directly relate to each of the dissertation projects, are provided below. Respectively, **Projects 1** and **2** aim to better understand and improve: The misuse of significance testing and the prevalence of *p*-value misconceptions (**section 3.1**), and researchers' ability to make sense of replicability in the wake of the replication crisis (**section 3.2**).

3.1. Dichotomous thinking & the misuse of statistics

Contemporary science has undoubtedly been marked by cultural and methodological reforms which have called into question the veracity of scientific research (e.g., Simmons, Nelson, and Simonsohn, 2011; Pashler & Wagenmakers, 2012). These reforms did not simply develop in light of uncertainty in conclusions, but as a consequence of more substantive reasons to doubt the reliability of the scientific method (mainly as it pertains to the use of null hypothesis significance testing (NHST); see e.g., Carver, 1978), and to question the norms of scientific conduct. Most notably, the Open Science movement has accelerated in recent years in alignment with growing concerns about the methodological and statistical rigour of scientific research, the prevalence of questionable research practices (QRPs) among scientists (see e.g., John, Loewenstein, & Prelec, 2012), and low replication rates of research findings, i.e. the 'replication crisis' (Open Science Collaboration (OSC), 2015). These and other concerns, concerning the culture of research (for further details, see **section 3.4** below), have led to widespread skepticism not only about the scientific integrity of research methods, but in turn about the credibility of published findings.

What is worth noting is that while reasons to question the trustworthiness of findings may have historically implicated cases of scientific misconduct (e.g., falsifying data or deliberate *p*-hacking), grounds for current distrust in research is also largely motivated by the *misuse* (rather

than the *abuse*) of statistical methods, which has been argued to stem from fundamental misunderstandings, markedly in the case of NHST (for an overview of criticisms, see Nickerson, 2000). Beyond theoretical criticisms of the method, such as its flawed logical structure (see e.g., Falk & Greenbaum, 1995), the misuse of significance testing among scientists was exemplified in the seminal work of Oakes (1986) in which he probed the baseline misconception rates of academic psychologists when it came to interpreting the meaning of *p*-values (see **Box 1** below for overview of *p*-value fallacies): His findings, which showed that 97% of surveyed researchers fell prey to at least 1 erroneous interpretation, raised a red flag in the research community, signalling a need to revisit the strengths and (perhaps more crucially) the limitations of the NHST approach as a model to infer conclusions and meaning from data. Since Oakes's initial work, the NHST controversy has not receded but arguably grown, despite the continued and dominant use of *p*-values in psychology (Cumming et al., 2007). In fact, the misuse of NHST is still very much an ongoing and real problem among psychology researchers, as evidenced in recent replications of high rates of *p*-value misconceptions (e.g., Haller & Krauss, 2002; Badenes-Ribera et al., 2015; Lyu, Peng, & Hu, 2018), as well as numerous cases of "interpretational overreach" (Spence & Stanley, 2018, p. 4) within the published literature (e.g., Vacha-Haase & Ness, 1999; Finch, Cumming, & Thomason, 2001; Hoekstra, Finch, Kiers, & Johnson, 2006) and in scholarly textbooks (Brewer, 1985).

This systematic pattern of documented error has been often attributed to the mechanical and formulaic use (i.e. "the null ritual"; Gigerenzer, 2004) of NHST, leading researchers to falsely assume "that the test of significance provides automaticity of inference" (Bakan, 1966, p. 423). Moreover, the emphasis on dichotomous thinking, i.e. "bucketing results into 'statistically significant' and 'statistically non-significant'" (Amrhein et al., 2019, p.306), has undermined the

complexity or nuanced-nature of research findings, and has led reform advocates (over 800 signatories just last year) to once again reject the use of arbitrary thresholds as signifying meaningful categorical differences: “In line with many others over the decades, we are calling for a stop to the use of *P* values in the conventional, dichotomous way – to decide whether a result refutes or supports a scientific hypothesis” (Amrhein et al., 2019, p. 306).

While there is no doubt that NHST has been heavily criticized for decades, being described at times by critics as “essential mindlessness in the conduct of research” (Bakan, 1966, p. 436), “a corrupt form of the scientific method” (Carver, 1978, p.378), or even “the most bone-headedly misguided procedure in the rote training of science students” (Rozeboom, 1997, p. 335), producing “nothing but an increase in pseudo-intellectual garbage” (Lakatos, 1978, p. 88) in place of scientific progress, there are several reasons that contribute to the appeal and persistent use of NHST (see Nickerson, 2000, for overview). Two key reasons summed up aptly by Nickerson (2000) include what was already touched upon above (see **section 2.1**), i.e. the “lack of understanding of the logic of NHST or confusion regarding conditional probabilities” (p. 246), as well as “the deep entrenchment of the approach within the field, as evidenced in the behavior of advisors, editors, and researchers” (p. 246). Taken together, these two reasons call for improvements in the widespread misuse of significance testing, and are the basis for the first dissertation project: **“Improving statistical inferences: Can a MOOC reduce statistical misconceptions?”**

While criticisms listed above may render such a task seemingly futile, there are two points work mentioning that offer optimism. The first is that while prevalent rates of *p*-value misconceptions have led some to claim that they are “impervious to correction” (Haller & Krauss, 2002, p.1), this tendency to assume that NHST misconceptions are deeply rooted in the

minds of researchers, or withstanding of proper instruction, appear to be largely a product of surveys like Oakes's (1986), rather than evidence that instructional interventions are ineffective. In fact, even if misconceptions that *are* resistant to change should produce widespread misconception rates, the converse does not necessarily logically follow: In other words, the observation that misconception rates are widespread does not entail that they stem necessarily from conceptions that cannot be refined or improved. This invalid reasoning form, i.e. *affirming the consequent,* is analogous to the *inverse probability* fallacy, a fallacy individuals commonly fall prey to when interpreting *p*-values (see **Box 1** below).

The other point concerns the work of Kalinowski, Fidler, and Cuming (2008), one of the very few studies that sought to investigate teaching interventions to improve *p*-value misunderstandings, specifically in terms of the aforementioned *inverse probability* fallacy. Authors appealed to the concept of "insight by comparison" (Haller & Krauss, 2002) as a method actively pinpoint contrasts between related but critically distinct concepts, in order to elicit commonalities as well as inconsistencies in reasoning. In particular, in one intervention, authors compared and contrasted NHST with Bayes' theorem in order to highlight that one cannot falsely assume the equivalence between $P(H_0 \mid D)$ and $P(D \mid H_0)$. In a second intervention, they explicitly drew attention to the common misapplication of the *modus tollens* argument form when categorical statements are replaced by probabilistic statements, i.e. how the validity of the (otherwise valid) deductive argument form breaks down when it is used to interpret the meaning of probabilities, such as *p*-values (for details, see **section 2.1**). As both interventions were found to be effective in improving the understanding of the *inverse probability* fallacy among students, Kalinowski and colleagues (2008) speculated that findings reflected the outcome of Gigerenzer's

concept of disrupting the null ritual; in other words, moving beyond the formulaic step-wise approach to more effectively teach NHST.

Box 1 – NHST Misconceptions: ***p*-value fallacies & correct definition**

- ***Inverse probability*** **fallacy** (Shaver, 1993; Kirk, 1996):
  The misconception that derives fundamentally from confusing the probability of the data, $P(D \mid H_0)$, with the probability of the theory, $P(H_0 \mid D)$, i.e. falsely assuming that one can draw conclusions about the probability of a theory or hypothesis, given sample data. This can manifest as the false belief that a *p*-value (e.g., $p = .01$) is equal to the probability that H_0 is true (i.e. 1%).

- ***Replication*** **fallacy** (Carver, 1978):
  The misconception that a *p*-value is directly related to the probability that an effect will replicate, i.e. falsely assuming that the *p*-value probability (e.g., $p = .01$, or 1%) can be taken as the complement of the replication probability (i.e. 1 - .01 = .99, or 99%).

- ***Effect size*** **fallacy** (Gliner, Vaske, & Morgan, 2001):
  The misconception that a *p*-value is directly related to the size of the effect, i.e. falsely assuming that a significant *p*-value necessarily entails a large effect size (or that a non-significant *p*-value necessarily entails a small effect size).

- ***Clinical or practical significance*** **fallacy** (Kirk, 1996):
  The misconception that a *p*-value is directly related to whether the effect in question is meaningful or not, i.e. falsely assuming that a significant *p*-value necessarily implies that the effect is practically important or clinically meaningful (or vice-versa for non-significant effects).

- **Correct definition**: The *p*-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true. The formal definition can be expressed as follows: $P(X \geq x \mid H_0)$ or $P(X \leq x \mid H_0)$, for right- versus left-tailed events, where X represents a random variable and x the observed event.

3.2. Making sense of 'failed' replications

This next section discusses the controversy surrounding the replication crisis, namely as it pertains to how researchers perceived and responded to the high rates of failed replications reported in the Open Science Collaboration (OSC; 2015) replication project. Here, the concept of treating effects categorically, characteristic of significance testing, is carried over to how the research community tends to treat the outcomes of replication effects: either a success or failure. Some speculations are then offered as to how this common way of thinking may have resulted in too-limited an understanding of how to judge replication research, leading some to be likely highly surprised by the low replication rates, and others to feel inclined to discredit the findings of the OSC (2015) project as evidence of a crisis in psychology. Specifically, misconceptions among researchers at the level of statistical indices, i.e. failure to fully comprehend the nature of frequentist concepts (*p*-values and CIs), may have contribute indirectly to misunderstandings in interpreting replicability at the level of cumulative research; eliciting, in turn, a crisis at the level of the scientific community.

Dichotomized framing is readily observed not only when researchers report on the outcome of a specific result (significant vs. non-significant), but also more generally when discussing the *existence* (vs. absence) of a phenomenon, the *truth* (vs. falsehood) of an effect, or the *success* (vs. failure) of replication attempts. Such an all-or-none way of thinking is consequently debilitating in face of contradictory results: "Because significance testing implies categorization of experimental outcomes, situations arise in which it is difficult to resolve controversies about whether a particular effect exists or not" (Hoekstra et al., 2006). The remark by Hoekstra and colleagues (2006), almost a decade prior, arguably forecasted how the routine dichotomizing of effects might have contributed to why the OSC (2015) replication study

elicited such a controversial response from the scientific community: Because the 100 original published studies being replicated were (up till then) broadly accepted or undisputed in the literature, it followed that when only a fraction of them successfully replicated (36% to 47% depending on the criterion), a 'replication crisis' broke out about whether the majority of studied effects in psychology were in fact true. Though there are many reasons for why one should not leap to such a generalized conclusion, it is nonetheless consistent with the popular intuition that if a pattern or result is repeatedly observed within a field, then it has a high probability of being real (Ioannidis, 2008; Simonsohn, Nelson, & Simmons, 2014), whereas failure to replicate undermines the likelihood that it exists. If you pair this intuition with the assumption that published effects in the literature are trustworthy, the wide-held false belief that significant *p*-values prove the existence of effects (i.e. *inverse probability* fallacy; Shaver, 1993; see **Box 1**), and the expectation that a significant finding has a necessarily high probability of replicating (i.e. *replication* fallacy; Carver, 1978; see **Box 1**), then it is additionally unsurprising that the OSC (2015) results became such a point of contention.

In response to the study, Gilbert and colleagues (2016) tried to dilute the impact of the study's results by identifying sources of error that were not accounted for (such as methodological infidelities and low power) so as to explain away and minimize the evidence of low replicability. In their comment, they stated:

> "If all 100 of the original studies examined by OSC had reported true effects, then sampling error alone should cause 5% of the replication studies to "fail" by producing results that fall outside the 95% confidence interval of the original study." (Gilbert, King, Pettigrew, & Wilson, 2016, p. 1037-a)

While Gilbert et al.'s (2016) criticism raised some fair points (like the issue of statistical power), it also demonstrated that misconceptions also exist when making sense of replication results. As illustrated in the excerpt above: Authors appealed to the widely held confidence-level misconception (CLM), which is the erroneous belief "that a 95% CI will on average capture 95% of replication means" (Cumming et al., 2004, p. 299). In reality, the capture percentage (CP) is closer to 83.4% on average, but can range significantly depending both on where the original study mean, as well as the replication mean, each fall in relation to the true population mean – two pieces of information which are by definition unknown (for details, see Cumming & Maillardet, 2006).

What is important to note here is that, while it is important to understand that "much scientific research is based on investigating *known unknowns*" (emphasis added; Logan, 2009, p. 712), it is just as important to simultaneously recognize the existence of *unknown unknowns* (M. Elson, personal communication, May 17, 2017), and to take those steps necessary to compensate for such sources of uncertainty. To elucidate in a few more words: While some factors in the research process are known to, say, affect sample estimates (for instance, in the case of meta-analysis, use of *p*-hacking techniques and publication bias are known to inflate meta-analytic effect size estimates), what is not known, for example, is which specific *p*-hacking techniques, and within which specific studies, these techniques were applied. Moreover, while one might knowingly attempt to correct for inflation of meta-analytic effect sizes (e.g., by trying to quantify the amount of inflation (*known unknown*) through the use of publication bias tests), such tests are incapable of distinguishing between inflation caused by *p*-hacking versus publication bias, or the interaction of both (*unknown unknowns*).

In the case of replication, a similar scenario is observed: On the one hand, when making sense of replication effects, some sources of uncertainty are known to exist, such as the two aforementioned sources of variation identified in the work of Cumming and Maillardet (2006) i.e. variation of the original mean around the population mean, and the variation of the replication mean around the population mean. Therefore, in order to effectively consider how two studies (original vs. replication) compare, authors should acknowledge, in a first step, the existence of these two sources of variance (*known unknowns*) and how they affect the CP of future replication estimates, which may help dispel the common CLM belief among researchers (roughly 46% prevalence; Cumming et al., 2004). Second, one must importantly recognize, however, that the very fact that the population parameter is unknown, means that knowing the extent to which either the original mean, or the replication mean, deviates from the population mean – and in turn which (original vs. replication mean) may better approximate the true mean – is simply not possible (*unknown unknowns*). As such, only by conducting a series of replication studies (direct and/or conceptual) can one begin to get a sense of where the population parameter might lie, and what might constitute an 'accepted' range of estimates for the specific effect in question. Without factoring in these layers of uncertainty and sampling variability (emphasized, resp., in **sections 2.1 & 2.2**), a researcher risks drawing inaccurate conclusions about the meaning of 'successful' or 'failed' replications, especially in face of contradictory findings.

While, as mentioned above, a series of independent replication studies (both direct and conceptual) is necessary to appreciate the totality of some *specific* effect – from establishing its basic existence to assessing the extent to which it is stable or justifiably generalizable (see LeBel, Berger, Campbell, & Loving, 2017) – research which looks at replicability across a set of *different* effects, as in the case of the OSC (2015) replication project, can also lend insight into

the nature of replicability. Beyond establishing general rates of replicability within a discipline, these replication projects have begun to explore the use of different criteria to judge *success* versus *degree* of replication, as well as predictors of replicability at the level of individual study features, such as *p*-values or effect sizes.

The second project of the dissertation, **"Exploring indices of repeated cross-validation as predictors of study replicability"**, builds directly off of this line of research, and seeks to explore the use of additional statistical indices as potential predictors to add to the cumulative research on the understanding of replicability of effects within psychology. What is interesting about this line of research is that, unlike Cumming & Maillardet's (2006) work mentioned above, which made use of simulations to explore replication and CPs, this second dissertation project involves the re-analysis of large high-powered real data sets to predict observed replication rates. While still in its infancy, opportunities to empirically explore the behaviour of real replication data sets, that have additionally (where possible) explicitly controlled for *known unknowns* (e.g., researcher degrees of freedom; see **section 3.4** below) have grown in recent years with the rise of meta-research (see **section 3.3** below).

3.3. Meta-research: The study of research itself

"Major disruptions are likely to happen in the way we pursue scientific investigation, and it is important to ensure that these disruptions are evidence based." (Ioannidis, 2018, p. 1)

Since the original replication project (OSC, 2015), efforts to estimate the replicability of scientific findings have risen in recent years, involving collective large-scale efforts like the Many Labs 1, 2, and 3 replication projects (Klein et al., 2014, 2018; Ebersole et al., 2016), the Experimental Economics Replication Project (Camerer et al., 2016), and most recently the Social

Sciences Replication Project (Camerer et al., 2018). Alongside these collective undertakings, notable advancements have also been made in the field of meta-research: "Meta-research is the study of research itself: its methods, reporting, reproducibility, evaluation, and incentives" (Ioannidis, 2018, p. 1).

Beyond investigating how research is being conducted, either at the level of individual studies or within sub-disciplines, a core objective of meta-research is also to gain a more global perspective on what makes for good scientific research, i.e. scientific investigations that more efficiently yield credible and useful research results (Ioannidis, 2018). In this way, "meta-research uses an interdisciplinary approach to study, promote, and defend robust science" (Ioannidis, 2018, p. 1). Therefore, central to effective meta-research is the ability to move between these levels of analysis to gain a comprehensive understanding of how different components of the research process interact, and in turn gain insight on how these interactions are unique or similar across different areas of research.

Not only is the overarching aim of the current dissertation in keeping with the objectives of meta-research, but given the complexities involved in the process of drawing accurate and meaningful inferences from statistical data, I would argue that this style of holistic interdisciplinary research approach, which attempts to bridge different elements involved in process of inference-making, is a promising framework through which to most comprehensively understand how researchers evaluate statistical evidence when drawing inferences from data, and how best to improve the art of effective statistical thinking and inference-making. While both dissertation projects may only offer two small disparate pieces of a much larger picture, hints about how misunderstandings may transcend or interact across different levels of analysis, might offer some initial empirical grounds upon which to speculate about more evidence-based

instructional supports to improve statistical inference-making (speculations elaborated in **General Discussion**).

Before moving onto the last section on how to foster better statistical thinking, one last component of the process of drawing statistical inferences which remains to be considered concerns the culture within research is carried out and inferences are routinely drawn (see **section 3.4** below). This can include research traditions, like NHST; however, because this was already covered above (see **section 3.1**), the following section will focus briefly on some potential sources of error or bias when interpreting statistical results, originating either at the level of the individual researcher (e.g., cognitive biases), or the community as a whole (e.g., incentive structures), which may contribute jointly to collective values or illusions regarding what should constitute informative or desirable effects.

3.4. Collective illusions & the culture of research

"Rituals call for collective illusions. Their function is to make the final product, a significant result, appear highly informative, and thereby justify the ritual." (Gigerenzer, 2004, p. 594)

The system of scientific inquiry, for any given discipline, is embedded within a culture of research, which includes, but is not limited to: traditions in research (either concerning theory or practice), accepted norms, as well as incentives. Beyond the ritual of NHST, psychology is rooted in a 'publish-or-perish' culture, a system that has been argued to pressure and reward "impact and productivity over quality and replicability" (Fanelli & Ioannidis, 2013, p. 15031). Such an incentive structure is likely to indirectly – if not directly – influence how researchers draw inferences from data, interpret statistical outcomes, and report results.

In discussing the social context of research, Klein et al. (2012) emphasize the existence of human sources of error that may unconsciously bias an experimental outcome toward success, consisting perhaps most importantly of "the twin (and sometimes incestuous) brother of demand characteristics, that is, *experimenter bias*" (emphasis added, p. 573). While their article was discussing bias at the level of individual participants and experimenters, their main idea has analogous implications at the level of the scientific community: Like participants, researchers are not "passive receptacles of stimuli" (Klein et al., 2012, p. 572). Given the high stakes to produce positive results, researchers are arguably less likely to approach their work with an impartial stance, and more prone to fall prey to confirmation bias, i.e. "the tendency to emphasize and believe experiences which support one's views and to ignore or discredit those which do not" (Mahoney, 1977, p. 161). Moreover, whether conscious or not, researchers are also likely to capitalize on researcher degrees of freedom to increase the likelihood of generating a significant result in order to produce a study that 'worked', and thus gain a competitive advantage. Such a culture of research reinforces carelessness (Martinson, Anderson, & De Vries, 2005) and indirectly promotes "'cutting corners' to achieve more interesting-looking results" (Ioannidis, 2018, p. 3). Self-serving behaviors of this nature have been referred to as "*p*-hacking" (Simonsohn et al., 2014, p. 534), "significance-chasing" (Ioannidis & Trikalinos, 2007, p. 247), or more generally as questionable research practices (QRPs), i.e. "the steroids of scientific competition" (John et al., 2012, p. 524). Not only have QRPs been described as 'normal misbehaviours' (De Vries, Anderson, Martinson, 2006), but their prevalent use was even judged as defensible among researchers (John et al., 2012).

The selective distortion of results is not limited to the treatment of the data (e.g., 'cleaning' or 'cooking' the data, or statistical adjustments), but also poignantly includes the

interpretation and reporting of results. These can for instance take the form of post-hoc predictions (e.g., HARKing: hypothesizing after the results are known; Kerr, 1998), inflated interpretations, or over-generalizations. One driving source of these distorted interpretations are presumably underlying misconceptions of what a statistical index means – e.g., conflating the meaning of statistical significance with practical relevance (i.e. *clinical or practical significance* fallacy; Kirk, 1996; see **Box 1**) or assuming that a significant *p*-value necessarily entails a large effect size (i.e. *effect size* fallacy; Gliner, Vaske, & Morgan, 2001; see **Box 1**). Similarly, over-generalizing is likely to be exacerbated by misunderstandings when inferring population characteristics from sample properties (see **section 2.2**), such as the *representativeness heuristic* and the *belief in the law of small numbers*, two misconceptions originating in misunderstandings of the law of large numbers (Tversky & Kahneman, 1971; Innabi & Jordan, 1999; Sotos et al., 2007). On the other hand, the pervasiveness of biased (e.g., confirmatory) interpretations has been recently argued to stem from a form of "cognitive opportunism" (Mercier & Sperber, 2017, p. 76); in other words, the exploitation of an evolved disposition in humans to apply post-hoc rationalization "to explain and justify *after the fact* the conclusions we have reached" (emphasis added, p. 112). Mercier and Sperber (2017) expressly distinguish between the concepts of 'inference' and 'reasoning', reserving the former for more spontaneous instances of information extraction, and the latter as the process of attending to reasons to accept a conclusion. Should this conclusion already be accepted inherently by the thinker in question, then reasoning in this context consists of a retrospective process of justification; arguably akin to *wishful thinking*.

Taken together, it is not surprising that, given the wide "latitude of rationalization" (John et al., 2012, p. 528) within psychology, that the published literature is saturated with positive findings (with odds of positive reporting bias approximately 5 times greater as compared to

'harder' sciences; Fanelli, 2010). In essence, not only does it appear to be "unacceptably easy to publish 'statistically significant' evidence consistent with *any* hypothesis" (Simmons et al., 2011, p.1), but negative (i.e. contradictory) findings are essentially non-existent. Consequently, it is even less shocking that researchers hold unrealistic expectations about the robustness or stability of effects (presumed to be true), and that they are particularly ill-equipped to deal with conflicting replication outcomes. Meta-research that seeks to understand how researchers approach data and make sense of results must also account for the suboptimal conditions within which statistical inferences are routinely drawn, including accepted norms, collective illusions, and cognitive biases.

Regarding the present dissertation, appreciating these potential sources of bias is directly relevant to the **Project 1** with regards to how interpretations are being made about *p*-values. Specifically, the main motivation in **Study 1** to include *p*-value items that also measure individuals' understanding about the meaning of non-significant outcomes (i.e. $p = .30$), in addition to items that probe the meaning of significant outcomes (i.e. $p = .001$), was to address this asymmetry in research norms that favour significant findings at the expense of publishing negative or null results. Consequently, not only are researchers faced with more opportunities to come across and familiarize themselves with significant outcomes as a result of the biased literature, but are also realistically provided with more instances to encounter over-generalized or inflated claims in reference to the meaning of a significant *p*-value, which may in turn introduce or reinforce existing misconceptions. Moreover, because to date, *p*-value misconception surveys have only assessed individuals' understanding of significant *p*-values, which may or may not be perceived or treated differently than less 'interesting' or 'desirable' outcomes, misconception rates may also be skewed, or at the very least establish an incomplete picture to work off of when

attempting to develop means to improve these misconceptions. As such, **Study 1** sought to first establish a slightly more nuanced understanding of how researchers interpret *p*-values by not limiting items to include interpretations about only an exclusive and unique fraction of the full set of possible *p*-values that researchers may observe.

4. FOSTERING EFFECTIVE STATISTICAL & META-SCIENTIFIC THINKING

Up to now, a great deal of emphasis has been placed upon discussing the complexities that surround the act of drawing accurate and meaningful statistical inferences from data, namely in the case of researchers within the discipline of psychology. Far from being a static process, inferential statistics involves dynamically assessing a set of considerations that extend well beyond the scope of the statistic or statistical index being interpreted, including theoretical constraints, design limitations, sampling properties, the nature of the data, etc. Moreover, inferring meaningful conclusions or implications of a statistical outcome requires inhibiting the tendency to fall victim to cognitive biases, such as the representativeness heuristic, the belief in the law of small numbers, confirmation bias or wishful thinking. Critically, the ability to recognize degrees of uncertainty that characterize inductive inferences should attenuate the odds of over-generalizing, making inflated claims, or post-hoc rationalizing when interpreting statistics. It follows that the art of effective statistical thinking importantly calls for researchers to move beyond a dichotomized way of framing their results, i.e. '*Is my finding significant (or not)?* or *'Did my study work (or fail)?'*, and instead ask thoughtful questions such as *'What would falsify my hypothesis?'* and *'Given the gap between data and theory, what may be justifiably*

inferred?'. In this way, effective statistical thinking, or the ability to engage in statistical reasoning, cannot be lessened to any one core skill (e.g., mathematics or logic).

With this in mind, when it comes to considering what kind of interventions or scaffolds might be best suited to foster effective statistical thinking, it might be sensible to consider improvements in understanding beyond the perspective of conceptual change models (e.g., novice-expert dichotomy), but also appeal to theoretical frameworks (e.g., problem-solving literature) that might elucidate how shifting one's pattern of thinking can fundamentally shape the way in which an individual frames a problem to be more apt to its solution; this relates back to how the ability to formulate salient and meaningful questions can be instrumental to effective statistical thinking (see **section 2.3**), and in turn the basis for well-founded inferences. Beyond lending alternative perspectives, the two bodies of literature (i.e. conceptual change vs. problem-solving) share some core aspects that map well onto the discussions provided thus far about the prevalence of statistical misconceptions, specifically with respect to the institutionalized nature of NHST in psychology.

First, both areas of research commonly appeal to the role of prior knowledge or previous experience when it comes to the predominant and/or persistent use of some given way of thinking. In terms of the conceptual change literature, this would be characterized by the existence of naïve or misguided conceptions among novices which stem from prior instruction, or develop from (and are likely reinforced by) real-world familiar experience. Only when these preconceptions are successfully replaced or refined, typically via formal instruction, do they become expert concepts (for details, see **section 4.1** below). In terms of the problem-solving literature, getting stuck within one frame of thinking and being unable to conceptualize a problem in a novel way, is typically a function of how an individual has habitually approached

the problem, shaped through past goals and experience. Being able to deviate from convention, whether it involves abandoning one's initial strategy for a more suitable one, or perceiving and applying a familiar tool/strategy in an unconventional way or to an unfamiliar context, is typically viewed as the process through which one overcomes *fixation* or *functional fixedness* (Maier, 1931; Duncker, 1945) and arrives at the problem solution (for details, see **section 4.2** below). Secondly, common across both bodies of literature is the notion of triggering a shift in thinking: In terms of conceptual change, *cognitive conflict* (Piaget, 1975; Posner ,Strike, Hewson, & Hertzog, 1982) is one mechanism through which naïve intuitions or conceptions develop into more expert concepts; when it comes to problem-solving, the act of *breaking frame* (or restructuring the problem space) is argued to facilitate effective problem reformulation (e.g., Maier, 1931; Dunker, 1945), affording in turn a vantage more suitable for problem solution.

When applied to the traditional use of *p*-values, which has been (and continues to be) the convention amid the research community, it is thus arguably among the majority of researchers the approach to inferential statistics most rooted in practical experience. As such, this "deep entrenchment of the [NHST] approach within the field" (Nickerson, 2000, p. 246), which has stood up to decades of criticism (see **section 3.1**), might explain why the field has not felt the impetus to move beyond a more naïve (i.e. dichotomous) use of *p*-values when drawing inferences. It may also explain why this overreliance on one strategy to approach statistical inference may persist even when ill-suited to answer the questions researchers are attempting to answer. In this respect, whether appealing to models of conceptual change, or traditions within the problem-solving literature, overcoming or breaking free from misguided habits when drawing inferences (i.e. fostering better statistical thinking) may entail a shift in tradition, involving fundamental changes in the way inferential statistics are routinely taught.

While I would argue that part of this change should move toward the instruction and exploration of multiple statistical techniques, or as Gigerenzer and colleagues would describe as instructing researchers the art of "opening the statistical toolbox and comparing tools" (Gigerenzer, Krauss, & Vitouch, 2004, p. 7) (see **section 4.3**), this is not to say that mere exposure to more statistical methods will necessarily undo pre-existing misconceptions, or prevent them from arising in the first place. Rather, comparing and contrasting why and how different statistical concepts and methods (e.g., frequentist vs. Bayesian) can be used to answer similar versus different questions, may play a critical role in teaching researchers to effectively appeal to different statistical tools and methods when drawing inferences from data (see **section 4.3** and **general discussion**). Prior to delving into this last idea, both bodies of literature mentioned above (i.e. conceptual change and problem-solving) are briefly discussed below (resp. **sections 4.1 & 4.2**).

4.1. Cognitive conflict: A model of conceptual change

Proponents of conceptual change theory (e.g., Posner et al., 1982) have commonly invoked the following learning framework when it comes to the acquisition of expertise: Conceptual change constitutes the key mechanism via which a novice, who initially holds naïve or misguided conceptions, replaces these incorrect intuitions with correct knowledge (i.e. expert concepts), typically as a result of formal instruction. Although the majority of the literature has pertained specifically to the domains of mathematical and scientific thinking, and has been investigated at the level of student learning (see Smith, diSessa, & Roschelle, 1993, for an overview), such a perspective has been also applied to the concept of statistical learning, "a view of learning as conceptual change [that] requires a shift from these 'naïve statistics' ideas, based

on everyday beliefs about probability and statistics, to accurate conception" (Finch & Cumming, 1998, p. 900).

While discussing the details of conceptual change theory goes beyond the scope of this work, it is worth briefly pointing out a set of assertions that have been described to characterize the misconception literature (see Smith et al., 1993 for details). First, misconceptions are framed in the context of a novice-expert *distinction*, whereby only through *replacement* can naïve conceptions be exchanged for expert concepts. Those who take a more constructivist approach to learning assert rather a novice-expert *continuum*, in which naïve intuitions play a productive role in learning; in other words, acquisition of expertise involves the gradual *refinement* (rather than *replacement*) of naïve beliefs into more advanced conceptions. It follows that the traditional perspective views misconceptions as necessarily interfering with learning, whereas the constructivist view reserves the possibility that misconceptions can be foundational in the development of expert reasoning. Both traditions agree on the assertion that learners are not blank slates when it comes to learning, and possess a set of prior beliefs (which may or may not be wholly accurate), stemming either from real-world experience or prior instruction. In keeping with these differences, some authors (e.g., Glaser & Bassok, 1989) distinguish between the ideas of valid versus invalid conceptions, i.e. *preconceptions* versus *misconceptions* (resp.), whereas others differentiate between "*misconceptions* – [i.e.] misunderstandings derived from instruction – from *alternative conceptions* – [i.e.] explanations students formulate as a result of their ordinary life experiences and bring with them to instruction (Driver & Easley, 1978)" (Guzzetti et al., 1993, p. 117). When it comes to statistical misconceptions, authors have typically not taken so strict a stance, taking the term 'misconception' to refer "to any sort of fallacies, misunderstandings, misuses, or misinterpretations of [statistical] concepts, provided that they

result in a documented systematic pattern of error" (Sotos et al., 2007, p. 99; Cohen, Smith, Chechile, Burns, & Tsai, 1996).

Another core tenet of the misconception literature is the notion of *cognitive conflict* (Piaget, 1975; Posner et al., 1982), which is viewed as conducive to conceptual change. Whether triggered by direct instruction, or by being confronted with plausible counterevidence or conceptual alternatives, such discordance should lead to system-level changes, prompting the learner to reformulate his or her prior intuitions to cohere with newly gained information. That said, misconceptions are often characterized as: "deep seated and resistant to change" (Clement, 1987, p. 3) despite proper instruction, being able to coexist in the mind of a learner alongside the correct understanding (Clement, 1982), and in even some cases defended by the learner (Smith et al., 1993). When a systematic pattern of error is observed among learners and/or experts, especially in spite of correct instruction, these types of misconceptions have been argued to be rooted in – and reinforced by – daily experiences in the real world (e.g., physics concepts): "Misconceptions that are persistent and resistant to change are likely to have especially broad and strong experiential foundations" (Smith et al., 1993, p. 152).

With regard to statistical misconceptions, most notably in the case of *p*-value misunderstandings, the systematic pattern of errors observed across students and researchers (e.g., Oakes, 1986; Haller & Krauss, 2002; Badenes-Ribera et al., 2015; Lyu, Peng, & Hu, 2018) have led some to describe these misconceptions as "impervious to correction" (Haller & Krauss, 2002, p. 1). While the authors made no explicit link to conceptual change models, such a description is nonetheless consistent with the notion of naïve conceptions that resist improvement, on account of ingrained false beliefs and assumptions. Moreover, the fact that seasoned researchers, and even methodology instructors, are not immune to misuse of

significance testing, reinforces the implication that formal instruction of frequentists statistics is not sufficient in triggering lasting conceptual change.

The trouble with these speculations however is that first and foremost, these studies only measured misconception rates at one time point and did not probe whether improvements could be facilitated by instruction. In other words: Even if misconceptions that *are* resistant to change should produce widespread misconception rates, this does not logically entail that widespread observation of misconception rates stem necessarily from preconceptions that cannot be refined or improved. This assumption cannot be presumed but must be empirically tested (i.e. **Project 1**). Moreover, instruction of NHST, which has been heavily and consistently criticized for being taught mechanically, i.e. as a rote ritual that encourages limited categorical thinking (e.g., Gigerenzer, 2004; see **section 3.1**), may itself contribute to the naïve or restricted understanding of *p*-values among researchers. Finally, even if some core misunderstandings, like the *inverse probability* fallacy (see **Box 1**, p. 23), have been argued to be based in a basic probabilistic misinterpretation (see Nickerson, 2000), which unjustifiably applies a deductive reasoning framework to an inductive inference scheme (see **section 2.1)**, this confusion between probabilistic versus categorical statements may be reinforced by the practice of treating statistical outcomes as categorical, a practice commonplace across researchers at all levels, and across subdisciplines. It is also thus imperative to empirically assess whether instruction, which deviates from the null ritual approach, can improve misconceptions (**Project 1**).

Unlike the misconception literature above, which would place more weight on preconceptions developed *outside* of formal instruction as a basis for the formation of misconceptions among novices, this perspective highlights how facets of how NHST is routinely taught and applied would in part explain why, despite decades of theoretical criticisms, the

misuse of *p*-values persists even among seasoned researchers. In this way, the notion of a novice-expert dichotomy or continuum is arguably not wholly applicable (or non-representative at best) in the case of NHST. Finally, because the NHST null ritual is so widespread, and alternative methods (e.g., Bayesian statistics) so sparse, it is difficult to begin to assess whether correct instruction or exposure to conceptual alternatives could effectively trigger conceptual change when it comes to interpreting the meaning of *p*-values (see idea of "insight by comparison", **section 3.1**) .

For these reasons, I will appeal additionally to the problem-solving literature (see **section 4.2** below) as an alternative or complementary framework which highlights the merits of being able to flexibly adapt one's perspective, especially when it comes to ill-defined problems, to suit the constraints of potentially similar yet distinct problems. Fostering this type of nuanced or dynamic outlook when approaching a problem space maps well onto the idea of effective statistical thinking, which necessitates appreciating and weighting multiple sources of evidence simultaneously in order draw well-founded inferences (see **section 2.3**).

4.2. Statistical tools & effective problem formulation

"The mere formulation of a problem is far more often essential than its solution."
(Einstein & Infeld, 1938, p. 83)

A key feature of effective statistical thinking, beyond the ability to optimally weight various pieces of evidence when drawing inferences, is the ability to ask salient questions (see **section 2.3**). This entails critically knowing which questions *can* and *cannot* be tackled, and to which degrees of certainty questions can be answered, given the choice of statistical method being used. Such an idea is not novel: Advocates of statistical reforms have referred to statistical

tests as *tools* (e.g., Mayo & Cox, 2006), the array of statistical tests as the *statistical toolbox*, and effective statistical thinking as "the art of choosing a proper tool for a given problem" (Gigerenzer, 2004, p. 588).

In the problem-solving literature, this can be construed as one component in the art of effective problem formulation. According to the search-inference framework (Baron, 2000), problem-solving involves moving from the initial *problem state* to the *goal state*, via a set of available operators, by selecting actions relevant to reducing the differences between these respective states; otherwise referred to as the *means-ends analysis* approach, and characteristic of the structure of Newell and Simon's (1976) General Problem Solver (GPS; 1976). Specifically, "a problem space is defined by mental representations of the initial problem situation, a set of relevant actions, and a goal" (Ohlsson, 2012, p. 118). As such, the way in which one construes his or her problem space, which is shaped not only by which problem characteristics are perceived as relevant, but also theoretical goals and epistemic aims, will necessarily play an operative role in how one engages in strategies in attempts to arrive at a solution. Given that there does not exist only one way to represent a problem, with some more apt than others, problem formulation is an integral part of effective problem-solving.

While the means-ends analysis framework is applicable for well-defined analytic problems, whose solutions can be obtained in a linear step-wise process, such a serial heuristic search through the problem space and possible solution paths will ultimately fail when the problem space itself is not appropriately or easily represented, as in the case with ill-defined problems:

> "In an ill-defined problem, uncertainty inheres not only in whether the goal will be reached but in how best to conceive the current state, goal state, and/or operators. The real problem, therefore, is how to develop a new problem formulation, transforming the ill-defined problem into a well-defined problem that can be solved." (DeYoung, Flanders, & Peterson, 2008, p. 279)

Given the complex and ill-defined nature of psychological phenomena, Newell and Simon's attempt at a general psychological theory (i.e. principle of heuristic search) failed. As Ohlsson (2012) points out, "there is no single problem solving mechanism, no universal strategy that people apply across all domains" (p. 117).

The Gestalt tradition, on the other hand, which offers a more holistic and nonlinear framework, has been adopted by scholars in the context of problem-solving (e.g., Köhler, 1970), where a problem's solution necessitates a fundamental shift in perspective or restructuring of the problem space. Such a need for problem reformulation, or *breaking frame*, is typical in the case of insight problems where one's initial conception of what is or is not relevant to a problem frame might lead to *fixation* (Maier, 1931) on components or characteristics of the problem that are not conducive to its solution. Insight, as such, has been described in terms of overcoming *functional fixedness* (Duncker, 1945), allowing for a novel representation of the situation to emerge. This process of breaking frame is commonly characterized as a response to encountering an impasse after attempted solution paths fail, and often "reflects the need to overcome the imperatives of past experience" (Knoblich, Ohlsson, Haider, & Rhenius, 1999, p. 1534). In line with the notion of *affordances* (Gibson, 1977), i.e. the perception of an object as a function of its action possibilities, the way in which one will perceive a problem and/or a potential tool for its solution will necessarily be shaped by the thinker's goals and past experiences. Likewise, a shift

in goals, or exposure to more tools (or more affordances of the same tool), should in theory foster a basis for more flexible and in turn more effective problem formulation.

With regard to the domain of psychological research, the problem of drawing inferences from data has been approached almost exclusively from the perspective of the step-wise NHST approach, whose misuse has also been attributed to misunderstandings about what the test of significance *can* and *cannot* afford in terms of statistical conclusions (e.g., Bakan, 1966; Shaver, 1993): "*p* values are simply not suitable for scientific inferences because they don't provide the information scientists *really want to know*" (Colling & Szucs, 2018, p. 7; see also Nickerson, 2000; Lindley, 2000). This reasoning has been invoked by Bayesian advocates (e.g., Wagenmakers, 2007; Dienes, 2011; Morey et al., 2016) to argue that Bayesian statistics can in fact solve this dilemma: Researchers are less interested in knowing the probability of their data (or more extreme data) under the null (i.e. $P(D \mid H_0)$ or the *p*-value), but rather more interested in the probability of their hypothesis given their (or more extreme) data (i.e. $P(H_1 \mid D)$, computed by Bayes' rule). Without implying that there is only one question worth posing, from this line of reasoning it can also be argued that: The greater the toolset provided to researchers, the higher the likelihood that it should contain the tool best suited to answer the question(s) being posed.

From the perspective of the problem-solving literature, the tradition of inferential statistics in psychology, to rely on the NHST formulaic approach, has likely stunted researchers in the following ways: First, it may have led researchers to commonly misframe the process of drawing statistical inferences as a well-defined problem, taking for granted as the goal state to produce a significant result. Moreover, such an overreliance on one approach is likely why researchers are not better equipped with a broader set of statistical tools through which to tackle research problems. Much in the same way that the GPS was criticized for assuming that a

universal problem-solving strategy exists, the overreliance on NHST in psychology can be criticized for naively presupposing that different research questions can be tackled in the same way, as well as fundamentally ignoring that what constitutes 'relevant' in one scenario will not necessarily be the case in the next:

> "We know but often forget that the problem of inductive inference has no single solution. There is no uniformly most powerful test, that is, no method that is best for every problem. Statistical theory has provided us with a toolbox with effective instruments, which require judgment about when it is right to use them." (Gigerenzer, 2004, p. 604)

An important caveat of course is that simply increasing exposure to more statistical methods will not necessarily produce more flexible or effective statistical thinking, nor necessarily protect against pre-existing versus newly acquired misconceptions for, respectively, familiar versus new concepts. Therefore, not just content, but also *how* instruction is carried out, should be jointly considered when attempting to develop interventions to improve statistical inference-making. Finally, before discussing this last point further (see **section 4.3** below), it is worth highlighting one final aspect of problem formulation that was already mentioned above, namely: that *how* one construes his or her problem space will include characteristics of the problem deemed relevant *as well as* situational constraints, such as theoretical goals and epistemic aims. Though more conceptual in nature, it could be argued that from a Gestalt perspective, pressures at the level of the scientific community (e.g., competitive incentives to publish and report impressive effects) might not only shape one's goal state (e.g., generate positive results), but accordingly influence which path of actions researchers perceive as relevant in attaining this goal (e.g., *p*-hacking, optional stopping, making over-inflated claims) versus those that may be viewed as unwanted hurdles to this goal (e.g., collecting large amounts of data,

reporting null results, being transparent about the meaning of one's findings). While this may just sound like an alternative way to restate what was already stressed in the section about research cultures (**section 3.4**) – i.e. that research pressures exist and should thus also be accounted for – the reason for framing it in these terms has to do with what might be surmised in terms of instructional implications, from the perspective of the problem-solving literature, specifically when it comes to accounting for constraints at this level. These ideas are now discussed below (**section 4.3**).

4.3. Breaking frame: Disrupting the null ritual

"It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail." (Maslow, 1966, p. 15-16)

Given the literature above, and taking into account the core criticisms that have loomed for decades around the misuse of statistics within research (e.g., NHST controversy), it behooves the scientific community to make active changes in the way in which researchers are taught to draw inferences from data. Current statistical reforms, especially those urging a shift away from dichotomous and categorical thinking in favour of a push toward use of alternative techniques, are accordance with Gigerenzer's notion of disrupting the null ritual, in order to cultivate a more meaningful – rather than mindless – approach to the use of statistics. Understanding from where sources of statistical errors originate, which factors and biases contribute to their existence, as well as how these misunderstandings can be improved, are just some questions that might be central to the success of these reforms. As such, more empirical research is warranted to effectively tackle these questions – a goal directly consistent with the projects of the current dissertation.

One common theme that has emerged throughout the text is the need to abandon, so to speak, the current tradition of over-relying on (significant) *p*-values, which to date seems to constitute for the vast majority of researchers both the main approach to – and the main goal of – inferential statistics. One idea raised, to alleviate this fixation on significance testing, was to introduce a broader range of statistical approaches, and their respective appropriate uses, to theoretically and practically equip researchers with a greater set of statistical tools and affordances when conducting research. Moreover, comparing and contrasting different methods, like the frequentist versus Bayesian models of inference, may be a means via which to trigger a sort of 'insight by comparison': By specifying how two statistical tools relate, or perhaps more critically how they fundamentally differ, may improve overall understandings of either or both methods. Finally, improving researchers' ability to ask questions when conducting research and interpreting data should also help guide individuals' in selecting and applying a statistical technique effectively. Importantly, the goal here would not be to "tell researchers what they want to know. Instead, we should teach them the possible questions they can ask" (for more on the Statistician's Fallacy; see Lakens, 2019, p. 5).

In keeping with these ideas, the first project of the dissertation **"Improving statistical inferences: Can a MOOC reduce statistical misconceptions?"** (Herrera-Bennett, Lakens, Heene, & Ufer) investigated the baseline misconception rates of *p*-values, confidence intervals (CIs), and Bayes factors (BFs) when it came to inferring the correct meaning of these statistical indices, as well as improvements gained as a result of participating in the 8-week online statistics course. **Study 1** extended past research (e.g., Oakes, 1986; Haller & Krauss, 2002) by challenging the assumption that systematic patterns of observed errors necessarily mean that misconceptions are resistant to change. Moreover, we measured whether differences in baseline

misconceptions rates arise when individuals are asked to draw interpretations about significant outcomes (i.e. when $p = .001$) versus non-significant outcomes (i.e. $p = .30$), which may point indirectly to the effects of cognitive biases (e.g., confirmation bias, wishful thinking) on the propensity to draw misguided conclusions. In **Study 2**, we investigated the effect of adding instructional material that explicitly pinpointed *p*-value misconceptions and helped train learners on how not to fall prey to common fallacies, to see whether learning can be further improved as well as maintained. While this project cannot speak to whether improvements in statistical understanding transfers beyond the scope of the course, and say into real research practice, it is necessary as a first step to gain an appreciation of how actively clarifying and disrupting prior intuitions does (or does not) elicit learning gains when it comes to inductive inferences, and/or what are the limits of these techniques. Additionally, though the sample was not exclusive to researchers in psychology, it is argued that since the course is taught from the perspective of psychology research, and that the majority of course users were pursuing academia at some level (e.g., BA, MA, prof), that study is thus relevant to gaining insights into the way researchers approach statistical inference-making. Lastly, while not central to the project's research questions, by investigating the effectiveness of a learning platform which covers a wide range of statistical concepts, including both frequentist and Bayesian approaches (taught back to back in weeks 1 and 2 of the MOOC, resp.), results may hint to the potential merit of equipping individuals with a broader range of statistical tools, and/or how deliberate comparison between approaches may improve the process of drawing statistical inferences from data.

Finally, in relation to the importance of asking questions when conducting research, the rise in meta-science in recent years has made it possible to ask questions today that previously were not possible to answer in practice. Specifically, high-powered replication projects can

provide initial insights into the nature of replicability in practical terms, beyond what theory has already taught us, such as: What are the observed rates of replicability in psychology and the social sciences (OSC, 2015; Camerer et al., 2018) versus economics (Camerer et al., 2016)? Given a set of direct replications of the same effect, what is the variation in replicability across different effects (Many Labs 1; Klein et al., 2014)? And how is variability in observed effect sizes affected by sample and setting characteristics (Many Labs 2; Ebersole et al., 2018) or participant pool quality (Many Labs 3; Ebersole et al., 2016)? As more of these works have continued to emerge, researchers have also tried to assess whether features of individual studies (e.g., sample size, *p*-value) reliably predict likelihood of a study to replicate. Being able to answer these questions will hopefully help dispel some of the misguided intuitions among researchers about replication and replicability rates, and/or provide hints as to why misunderstandings of a concept at one level (e.g., meaning of a single *p*-value) may relate to misunderstandings at a higher level (e.g., meaning of a distribution of *p*-values).

With these goals in mind, the second dissertation project **"Exploring indices of repeated k-fold cross-validation as predictors of study replicability"** (Herrera-Bennett, Ong, & Heene) investigates whether resampling techniques, namely cross-validation (CV), can be used to gain a deeper understanding about the relationship between individual study features (i.e. estimates of model fit and error: average *R*-squared, RMSE, MAE) and its likelihood to replicate. Through the re-analysis of the Camerer et al. (2018) replication project, this project on the one hand aimed to extend past replication research regarding the investigation of study features as potential predictors of replication success. On the other hand, the project also sought to test the theoretical assumption that cross-validation can serve as a model validation technique in practice if used to predict replication rates of real data. While CV theoretically provides a measure of

how the statistical model will generalize to and perform within an independent data set, rarely is this assumption able to be carried out in practice in a strict sense. Therefore, this project may provide some novel insights about the practical implications of CV which until now have received little to no investigation beyond simulation studies. Finally, in accordance with Thompson's (1995) remark about bootstrapping, CV analyses are likely to "capitalize during resampling on the commonalities inherent in a given sample in hand" (p. 95). In other words, because both the training and test samples do not capture variation associated with different experimental implementations across replication attempts, then the idea behind CV as providing an 'out-of-sample' generalization error estimate – while perhaps theoretically sound – might be an optimistic one. To this end, the project also speaks to the importance of empirically testing statistical assumptions, especially those that may be used to support claims about the validity of a model, such that as researchers, we continue to learn about the strengths and limitations of different statistical approaches.

Project 1

Improving statistical inferences: Can a MOOC reduce statistical misconceptions?

Arianne Herrera-Bennett[1], Daniël Lakens[2], Moritz Heene[1], Stefan Ufer[3]

[1]*Department of Psychology, Ludwig Maximilian University of Munich, Germany*

[2]*Department of Industrial Engineering & Innovation Sciences, Human Technology Interaction, Eindhoven University of Technology, Netherlands*

[1]*Department of Mathematics, Ludwig Maximilian University of Munich, Germany*

Abstract

Perseverant rates of statistical misconceptions have raised doubts about the methodological and statistical rigour in psychology and surrounding domains, leading some to claim that intrinsic misconceptions are "impervious to correction" (Haller & Krauss, 2002, p. 1). Surprisingly, little work to date has empirically investigated the extent to which statistical misconceptions can be improved reliably within individuals, nor whether these improvements are transient or maintained. Study 1 (N = 2,320) evaluated baseline misconception rates of p-value, confidence intervals, and Bayes factor interpretations among online learners, as well as rates of improvement in accuracy on these items across an 8-week MOOC. Given that the MOOC was not designed to specifically target statistical misconceptions, but rather to cover core concepts, Study 2 (N = 1,301, *preliminary data*) investigated the added effects on improvement rates for *p*-value interpretations by having an experimental group (n = 649) complete an additional assignment in week 1 of the MOOC, developed to explicitly cover and clarify common *p*-value misconceptions, as compared to controls (n = 652). Both Studies 1 and 2 demonstrated statistically significant improvements in accuracy rates, across all three concepts, for both immediate learning (at *post-test1*), as well as retained learning until week 8 (*post-test2*). Study 2 provided preliminary evidence to suggest that learners who underwent explicit training and clarification on common *p*-value misconceptions demonstrated greater improvements in learning, i.e. fell prey to fewer *p*-value misconceptions, as compared to controls. Taken together, the current work challenges the idea that statistical misconceptions are not flexible to change, and provides empirical evidence for the effectiveness of enhanced statistical training tools and online learning platforms. Moreover, these studies speak importantly to the role of explicit clarification when it comes to teaching statistical concepts and how to correctly draw statistical inferences.

Abstract word count = 291

Keywords: MOOC, *p*-values, misconceptions, statistical inferences

Improving statistical inferences:
Can a MOOC reduce statistical misconceptions?

Introduction

Doubts have been raised about the methodological rigour of psychology and the social sciences, evidenced by the documented errors in the scientific literature (Casadevall & Fang, 2012). A common concern that is assumed to contribute to these errors is misconceptions when drawing statistical inferences. The misuse of statistics and pervasiveness of misconceptions among researchers has been argued for decades to stem from fundamental misunderstandings, markedly in the case of null hypothesis significance testing (NHST; for an overview of criticisms, see Nickerson, 2000). In fact, beyond being described as "numerous and repetitive" (Badenes-Ribera et al., 2015, p.290), some authors have deemed *p*-value misinterpretations "impervious to correction" (Haller & Krauss, 2002, p. 1).

The trouble with such absolute claims is that although misconception rates appear to be relatively stable over time, this does not mean that misconceptions cannot be corrected *given clarification or training*. Yet surprisingly, little work to date has empirically investigated the extent to which statistical misconceptions can be improved reliably within individuals through training, nor whether these improvements are transient or maintained. The current research aims to address these questions.

Despite the reform toward the use of alternative or complementary statistical techniques, such as confidence intervals (CIs) and effect sizes (e.g., Bakan, 1966; Falk & Greenbaum, 1995; Cohen, 1990; Rozeboom, 1997; Wasserstein & Lazar, 2016), *p*-values continue to dominate the published literature in psychology (approx. 97% of articles across 10 leading journals; Cumming et al., 2007). Furthermore, "mindless statistics are not limited to *p*-values" (Lakens, 2019, p. 9): Alternative approaches such as CIs and Bayesian inference have their own possible misinterpretations (for overviews, see Greenland et al., 2016; Morey, Hoekstra, Rouder, Lee, &

Wagenmakers, 2016). As such, simply replacing p-values with other statistical tools is unlikely to resolve the problem that researchers have misconceptions about the statistical inferences they make (Lakens, 2019). Instead, we have to develop ways to improve the way people interpret statistics.

To this end, Study 1 investigated misconception rates among learners in the context of an online 8-week Coursera MOOC (massive open online course) aimed at the improving the process of drawing statistical inferences. To extend previous work, our sample was not limited to psychologists, but included learners across various disciplines. Moreover, while previous research has focused on inferences about statistically significant *p*-values (i.e. $p = .01$ or .001), our work also examines interpretations of CIs, Bayes Factors (BFs), and non-significant *p*-values (i.e. $p=.30$). Given that the MOOC was not initially designed to specifically educate learners about statistical misconceptions, but rather to cover core concepts as a whole, Study 2 investigated whether instructional material which explicitly targets and aims to clarify misconceptions about *p*-value can further bolster improvements in learning. Are misconceptions about p-values really impervious to correction, or can good educational material provide a source of hope?

Statistical Misconceptions: Frequencies and Common Fallacies

Since the seminal work of Oakes (1986), which found that 97% of academic psychologists fell prey to at least one erroneous *p*-value interpretation, misconception rates have appeared to remain relatively stable: Similar frequencies were observed among German (N=113; Haller & Krauss, 2002), Spanish (N=418; Badenes-Ribera et al., 2015), and Chinese (N=246; Lyu, Peng, & Hu, 2018) samples, and misconceptions are not limited to the field of psychology (N = 221 communication researchers; Rinke & Schneider, 2018). Although accuracy rates varied across academic status (Haller & Krauss, 2002), and qualification (Lyu et al., 2018), differences were only marginal. In fact, even methodology researchers and instructors, who made comparatively fewer incorrect interpretations (Haller & Krauss, 2002; Badenes-Ribera et al., 2015), were not immune to general misunderstandings (see also Lecoutre, Poitevineau, & Lecoutre, 2003), lending strength to the assumption that *p*-value misconceptions withstand proper instruction or training. These findings are evidence for the need to improve statistical education.

The current research strives to tackle this need, specifically drawing upon the work by Badenes-Ribera and colleagues (2015), who identified the rates of four specific *p*-value fallacies,

namely the *inverse probability* (Shaver, 1993; Kirk, 1996), *replication* (Carver, 1978; Fidler, 2005; Kline, 2013), *effect size* (Gliner, Vaske, & Morgan, 2001), and *clinical or practical significance* (Kirk, 1996) fallacies. The *inverse probability* misconception derives fundamentally from falsely assuming that one can draw conclusions about the probability of a theory or hypothesis, given sample data. Instead, the correct interpretation entails the converse: The *p*-value determines the probability of observing data (or more extreme results), *contingent* on having assumed the null-hypothesis is true. The formal definition can be expressed as follows: $P(X \geq x \mid H_0)$ or $P(X \leq x \mid H_0)$, for right- versus left-tailed events, where X represents a random variable and x the observed event. The *replication*, *effect size* and *clinical or practical significance* fallacies – i.e. respectively, the false assumptions that *p*-values are directly related to replication rates, effect size, and practical significance – are examples of inflated interpretations or over-generalizations when it comes to communicating the meaning and implications of a significant *p*-value. Specifically, in the case of the *replication* fallacy, individuals falsely take the *p*-value probability (e.g., $p = .03$, or 3%) as the complement of the replication probability (i.e. 1 - .03 = .97, or 97%). For the *effect size* fallacy, a significant *p*-value is falsely assumed to necessarily entail a large effect size (or a non-significant *p*-value a small effect size); in the case of the *clinical or practical significance* fallacy, a finding that is statistically significance is conflated with the idea that it is practically important or clinically meaningful.

Badenes-Ribera et al. (2015) demonstrated that academic psychologists were particularly prone to the *inverse probability* fallacy (93.8% error rate), a fallacy which has been deemed "the most common, and arguably the most damaging, misinterpretation of *p* value (Oakes, 1986)" (Kalinowski, Fidler, & Cumming, 2008). In contrast, the *replication*, *effect size* and *clinical or practical significance* fallacies incurred relatively lower error rates (resp. 34.7%, 13.2%, & 35.2%). Interestingly, nearly 50 years prior, Bakan (1966) referred to such assumptions as a form of researcher bias; that is, an intrinsic misattribution about *p*-value characteristics stemming from placing too much weight on the role of null hypothesis significance testing (NHST). In other words, when statistical tests or outcomes assume the "burden of scientific inference" (Bakan, 1966, p. 423), researchers "tend to credit the test of significance with properties it does not have, […] and overlook characteristics that it does have" (Bakan, 1966, p. 423-428). Similar impressions were shared by other critics of the NHST approach, considering the approach as harmful "because such tests do not provide the information that many researchers assume they do" (Shaver, 1993, p.294), in turn

allowing a *p*-value to be "interpreted to mean something it is not" (Carver, 1978, p. 392). This framework can be directly applied to our understanding of *p*-value fallacies, namely wherein the process of quantifying the outcome or implications of a study, the measure of statistical significance is being confounded with measures of impact (e.g., *effect size*) or meaningfulness (i.e. *clinical or practical significance*).

Misconceptions have also been identified with regard to CI interpretations, including the false belief that overlapping CIs necessarily imply a statistically non-significant mean difference (Belia et al., 2005), that CIs reflect the probability of containing the true population value (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), as well as the confidence-level misconception (CLM), i.e. the (erroneous) belief "that a 95% CI will on average capture 95% of replication means" (Cumming, Williams, & Fidler, 2004, p. 299; Cumming & Maillardet, 2006). Prevalent misconception rates have been observed among undergraduates and graduate students, as well as researchers (see Hoekstra et al., 2014, Lyu et al., 2018; Rinke & Schneider, 2018). In turn, methodological reforms, such as calls for improved training in probability and statistics, have been proposed to "enhance conceptual rigour and reduce the likelihood of a false conclusion" (Casadevall & Fang, 2012, p. 894).

Improving Statistical Inferences: Intervention Research

To date, despite the many articles lamenting how common misconceptions about statistical inferences are, few studies have aimed to examine interventions that could reduce misconceptions, particularly when it comes to *p*-values. Two studies, however, provide some initial insights, both geared toward overcoming students' tendency to conflate P(D|T) with P(T|D). The first is the work by Falk and Greenbaum (1995) who investigated, among a sample of university students (N = 53), the effect of warning students about this specific misconception prior to assessing their understanding of statistical significance. Specifically, students were asked to read Bakan's (1966) paper as part of one of their courses before being tested on the meaning of a significant *p*-value item. Results indicated that only 13.2% of individuals endorsed the correct interpretation, leading authors to conclude that their findings corroborated Oakes's (1986) misconception rates. In particular, authors attribute high misconception rates to subjects' inability to overcome "the (erroneous) belief that one has rendered the null hypothesis improbable by obtaining a significant result," (Falk & Greenbaum, p. 76), a misconception the authors label "the illusion of probabilistic proof by

contradiction" (p. 76). In response to the observation that vast majority (79.2%) of responses endorsed the interpretation of a significant *p*-value as "we showed that H_0 is improbable" (p. 85), the authors explain that:

> A significant test result virtually *asks* to be interpreted that way. The test was invented from the beginning to *reject* the null hypothesis, namely to show that it is unlikely to be true. People manage therefore to skillfully skip over the fact that the test computes the probability of (at least as extreme) *results, given* H_0, instead of giving the probability of H_0 conditioned on the results. (p. 84)

Such criticism aligns with Gigerenzer et al.'s (2004) concept of 'the null ritual', i.e. researchers' habit of engaging "in a statistical ritual rather than statistical thinking" (p. 2). In other words, rather than considering NHST as suitable to answering one's research question, use of significance testing is taken at face value as the default 'all-purpose' approach, whose logic as a result has been collapsed into to a set of mindless steps: Let H_0 = "no mean difference" or "zero correlation"; set alpha = 5%; if $p < \alpha$, then reject H_0, and accept H_1.

Disrupting such a ritual, by for instance contrasting the logic of NHST with the Bayesian approach, is argued to be a means via which a learner may more readily distinguish between what a researcher *hopes* to find, i.e. $P(H_1 \mid D)$, versus what a *p*-value actually indicates, i.e. $P(D \mid H_0)$. Pointing out this contrast, or 'inverse', is what Haller and Krauss (2002) refer to as the "insight by comparison" (p. 11), and what Kalinowski et al. (2008) used as one of their teaching interventions in a sample of undergraduate students (N = 120) to overcome the *inverse probability* fallacy. Authors found that either by contrasting significance testing with Bayes' theorem, or using counterexamples to highlight how probabilistic statements can be rendered invalid, students' improved their understanding from a baseline of 4.0 misconceptions (max. of 6.0) to only 2.0 post-intervention, and 2.7 at follow-up (5 weeks later). Such research provides initial insights on how explicit training on misconceptions, via for instance actively contrasting statistical approaches (frequentist vs. Bayes), may serve as instrumental in fostering a better understanding about the meaning of *p*-values. Additional research should not only aim to replicate such improvements, but also tackle other misconceptions that commonly exist among learners when drawing conclusions from data.

Study 1

Study 1 aimed at investigating the baseline misconception rates among online learners, as well as improvements, across the 8-week MOOC, when it came to drawing inferences about the meaning of *p*-values, CIs, and BFs. Because, to date, research has focused exclusively on how individuals draw inferences about statistically significant outcomes (i.e. $p = .01$ or .001), Study 1 also examined misconception rates for non-significant outcomes (i.e. $p = .30$), comparing baseline accuracy rates for *p*-value items in scale versions 1 ($p = .001$) and 2 ($p = .30$), which target the same misconceptions but are phrased to be consistent with significant versus non-significant outcomes.

Procedure

Building on previous work, a 14-item True/False scale was developed, targeting concepts of *p*-values (8 items), BFs (3 items), and CIs (3 items). Before arriving at this final 14-item scale, initial piloting ($N = 216$) was first carried out on an 11-item *p*-value scale adapted from previous scales (i.e. Haller & Krauss, 2002; Badenes-Ribera et al., 2015). Specifically, four versions of the 11-item scale were constructed to test how comparable items performed when interpretations concerned a significant (i.e. $p = .001$; versions 1 & 2) versus non-significant *p*-value (i.e. $p = .30$; versions 3 & 4); versions were also counterbalanced for negative-phrasing. After piloting, one item was dropped due to poor item discrimination across all four versions (i.e. corrected item-total correlations all < .243). The resulting 10-item scales yielded the following levels of internal consistency (Cronbach's α): version 1 ($n = 52$; $\alpha = .74$), version 2 ($n = 64$; $\alpha = .67$), version 3 ($n = 45$; $\alpha = .60$), and version 4 ($n = 55$; $\alpha = .69$); when items were collapsed across versions ($N = 216$), $\alpha = .68$ (as the scale was not assumed to be unidimensional in nature, but rather tapping into different types of *p*-value misunderstandings, scale consistency was considered acceptable). Additionally, pilot results revealed that two items had systematically worse item discrimination when reverse-phrased, and thus were not reverse items in the final 14-item scale (namely, PV2 & PV8). In the end, two versions (details below) of the finalized 14-item scale were used for Study 1: *p*-value items were reduced to 8 in total (dropping a couple redundant items), and 3 items each were added for the BF and CI concepts (see **Appendix A & B**).

Some important differences from the original scale administrations, aside from edits in item wording (e.g., reverse-phrasing in some cases), included prefacing items with the instruction that

"Several or none of the statements may be correct" (as was done in the Haller & Krauss, 2002, study), as well as including an *"I don't know"* response option (this is because as some concepts were expected to be completely new to some learners, the added response option was expected to discourage guessing). Additionally, two versions of the scale provided alternate wordings for the same misconceptions (see e.g., items CI1 or BF1) as well as items measuring both statistically significant (p = .001) and non-significant (p = .30) outcomes (respectively, **Appendix A** and **B**); note that *p*-value items across both versions differ only in regard to statistical outcome (i.e. p = .001 vs. .30) and not in terms of phrasing (i.e. reversed or not). Both versions were implemented (item order fixed-randomized and scale versions counterbalanced across participants) in the form of six "Pop Quizzes" (PQs) within the 8-week MOOC, namely the course *"Improving your statistical inferences"* taught by Daniël Lakens. Pre-/post-test design with three measurement periods served as proxies of: i) prior knowledge (*pretest*, i.e. PQ1; administered week 1), ii) immediate improvement (*post-test1*, PQs2-5; weeks 1-5, directly after the relevant lecture), and iii) retained learning (*post-test2*, i.e. PQ6; week 8). Post-test1 items were staggered across weeks 1 to 5 in order to occur immediately after the relevant module whose content pertained to the concepts in question. Questions were clustered into four subsets: subsets 1 and 2 (*p*-value items), subset 3 (BF items), and subset 4 (CI items) (see **Appendix A & B** for details). Demographics (e.g., self-rated statistical expertise level, level of education obtained) as well as confidence ratings of responses (7-point Likert scale ranging from 1 *"Not at all confident"* to 7 *"Very confident"*) were voluntarily requested.

Exclusion Criteria

Due to the flexible nature of the online course, participants could complete the pop quizzess more than once, and move between modules at their own pace. Therefore, data only included subjects' first response attempts, and excluded any users that did not complete measurements in the expected order (e.g., if participants completed *post-test* before *pretest*). Response latencies (i.e. lag between any 2 measurement occasions, measured in hours) were corrected for skewedness, excluding outliers (+/- 3 *median absolute deviations*; Leys, Ley, Klein, Bernard, & Licata, 2013); resulting asymmetry fell within acceptable limits (i.e. between -2 and +2; George & Mallery, 2010; see **Appendix E**).

Results (Study 1)

Sample

Data collection ran for 1 year (Aug 2017 – Aug 2018). Total number of MOOC learners at *pretest* was N = 2,320; of those learners who responded voluntary to the demographic questionnaire (n = 611), 57.8% were male (39% female; 3.3% opting not to specify), with a mean age of 37.93 years (SD = 10.77). Thirty-four percent of users reported English as their native language, and 82.0% of having previously taken a statistics course before participating in the MOOC. When asked to rate their level of statistical knowledge and understanding, 38.9% rated themselves as beginners, 54.2% intermediate, and 6.9% advanced. Regarding academic experience, about a third of individuals held a bachelor's degree or lower (29.7% with high school diploma or bachelor's degree), roughly half had completed graduate-level training (51.6% with Master's or PhD degrees), and the rest held post-graduate degrees (14.4% post-doctoral degree, 4.3% professorship).

Response Trends

In keeping with typically high MOOC dropout rates (average completion rates approx. 5%; for review, see Feng, Tang, & Liu, 2019), marked response attrition was observed across the six quizzes, with total number of learners at PQ1 (n_1 = 1,915) dropping to n_2 = 1,045 (PQ2), n_3 = 621 (PQ3), n_4 = 421 (PQ4), n_5 = 371 (PQ5), and n_6 = 276 (PQ6) respondents. Sample sizes for pop quizzes and reported mean scores exclude for cases with missing data on any of the respective quiz items (see **Figure 1** & **Table 1**). Notably larger proportions of "I don't know" responses were observed at *pretest* (27% at PQ1) as compared to *post-tests* (approx. 1% - 3% of responses across PQs 2-6). Mean confidence ratings remained relatively stable across all pop quizzes (ranging from 4.04 – 5.80; see **Table 1**).

Baseline Misconception Rates

Baseline rates were broadly consistent with past research insofar as general trends across *p*-value fallacies: The *inverse probability* fallacy was among the more difficult fallacies, with the *replication* fallacy displaying poorest rates, whereas the *effect size* fallacy had greatest accuracy. With that said, because the original studies only provided respondents with "True" or "False" options, whereas the current study also allowed for "I don't know" responses, two sets of accuracy

rates were computed and reported: proportion of correct responses as a function of all responses (i.e. "I don't know" coded as incorrect; see **Fig. 2A&C**), versus only attempted responses (i.e. "I don't know" responses omitted; **Fig. 2B&D**). CIs and BFs yielded greater were observed as relatively less familiar concepts across users, i.e. yielding greater proportions of "I don't know" responses. What is worth noting is relative difficulty levels across the different statistical concepts was fairly consistent given the different scoring computations (i.e. whether "I don't know" responses were coded as incorrect vs. omitted). The one exception was the BF items: Rates tended to show that among those who attempted to answer the BF items (i.e. did not opt for the "I don't know" option), accuracy was markedly higher. Specifically, when "I don't know" responses were coded as incorrect, the 3 BF items yielded accuracy rates of 14% (BF1), 47% (BF2), and 27% (BF3); in contrast, when only True and False options were compared, accuracy was 43%, 86%, and 64%, respectively.

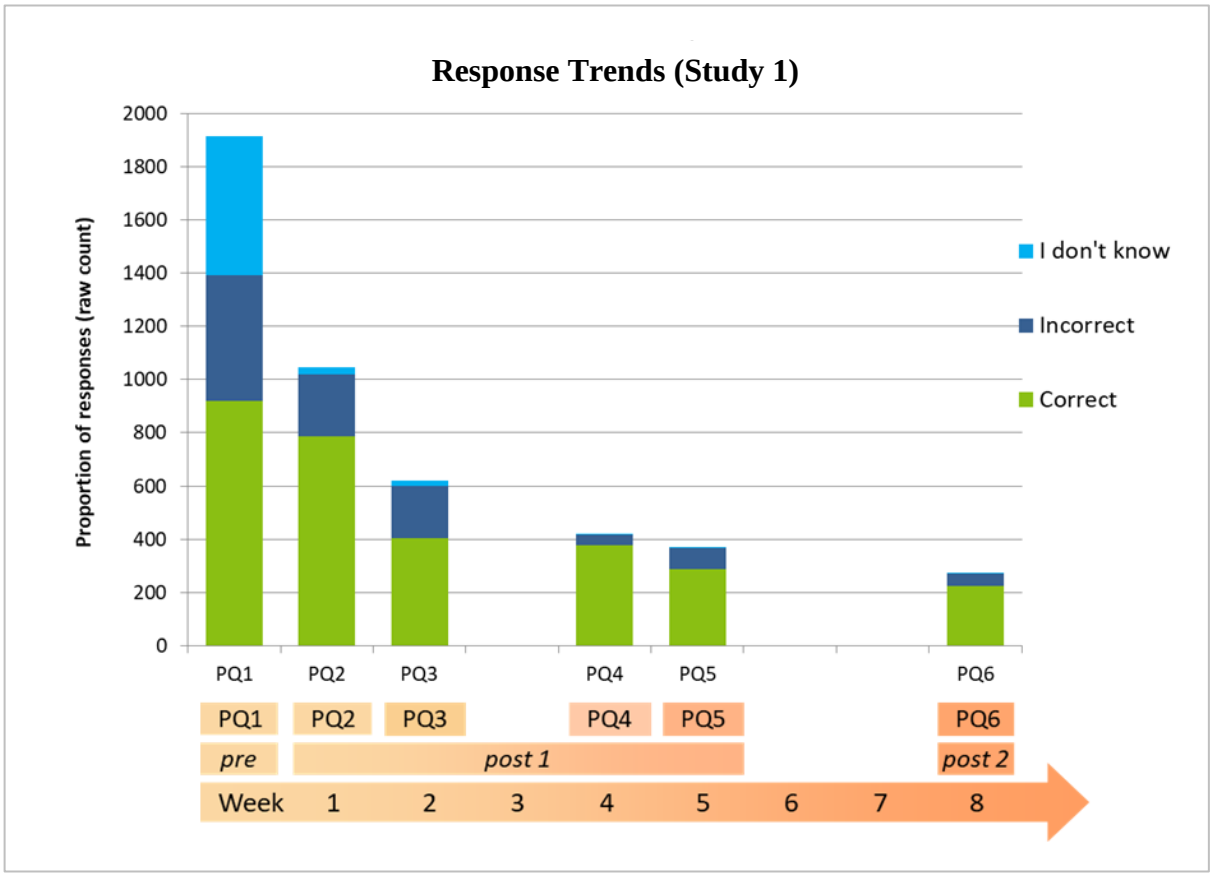

Figure 1 – Response trends (Study 1). Proportion of responses (correct, incorrect, "I don't know") across the 8-week MOOC timeline, i.e. from Pop Quizzes (PQs) 1 to 6. *Pre* = pretest (PQ1), *post1* = first post-test (summed across PQ2 – PQ5), *post2* = second post-test (PQ6).

Table 1

Response trends & confidence levels (Study 1). Accuracy rates (i.e. mean proportion of correct, incorrect, & "I don't know" responses), and mean confidence ratings (CRs), across pop quizzes (PQs) 1 to 6. Items covered in each PQ were clustered into four subsets: subsets 1 and 2 (*p*-values), subset 3 (BFs), and subset 4 (CIs).

	PQ1	**PQ2**	**PQ3**	**PQ4**	**PQ5**	**PQ6**
Items covered	Subsets 1 – 4	Subset 1	Subset 3	Subset 2	Subset 4	Subsets 1 – 4
N	1,915	1,045	621	421	371	276
Correct (%)	48.06	75.26	64.93	89.73	77.43	80.84
Incorrect (%)	24.67	22.22	31.70	9.33	21.53	17.90
I don't know (%)	27.26	2.54	3.33	0.97	0.97	1.26
mean CR *(7-pt Likert)*	4.04	5.06	4.74	5.80	5.35	5.17

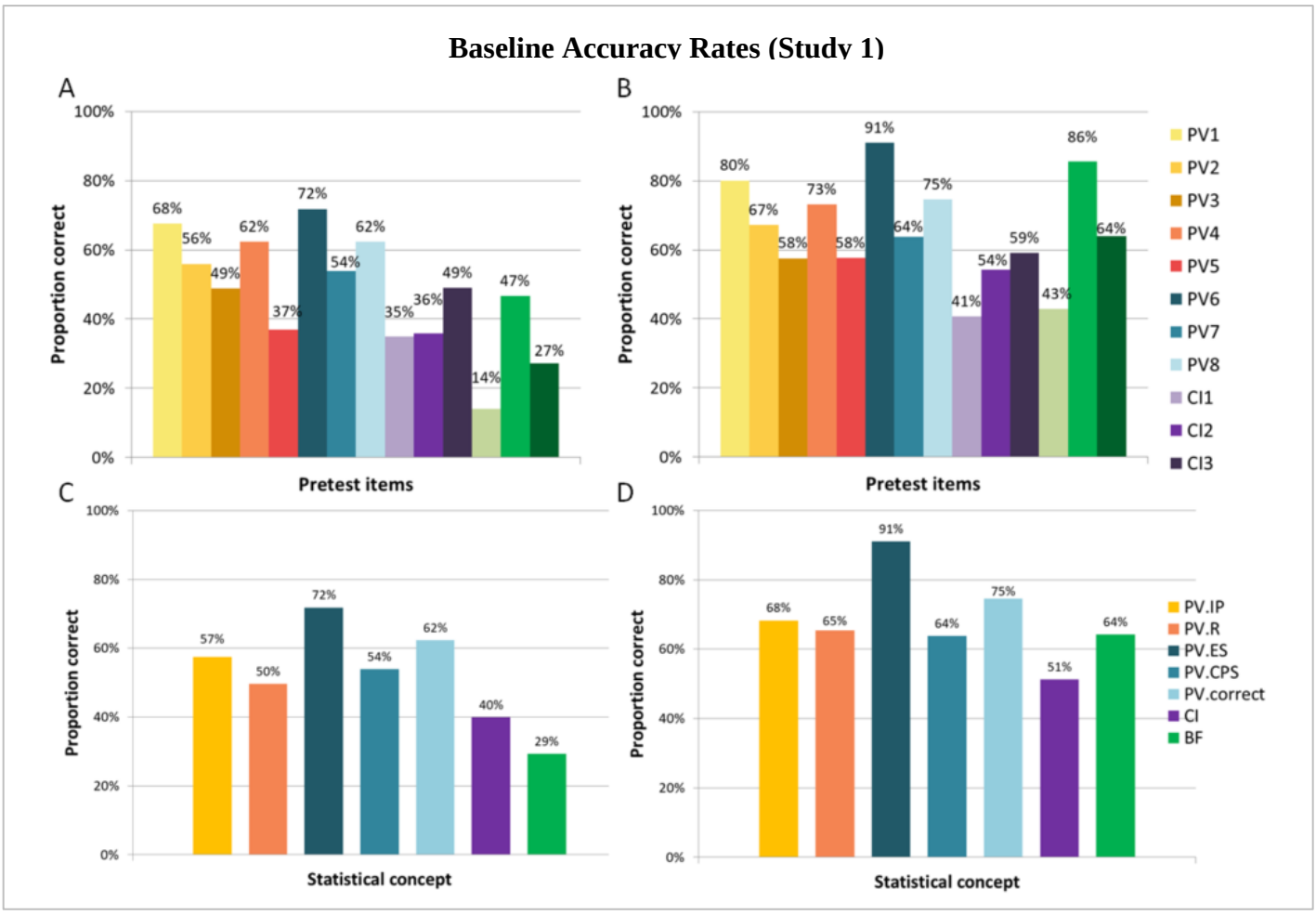

Figure 2 – Baseline accuracy rates (Study 1). Accuracy rates across the 14 individual items (**Figures 2A & 2B**), and averaged across statistical concepts (**Figures 2C & 2D**). Accuracy rates are computed in two manners. **2A & 2C:** As the proportion of correct responses given all provided responses (i.e. "I don't know" responses coded as incorrect); **2B & 2D:** As the proportion of all 'attempted' responses (i.e. "I don't know" responses omitted). Individual items consist of 8 *p*-value items (PV1 to PV8), 3 confidence interval items (CI1 to CI3), and 3 Bayes factor items (BF1 to BF3); statistical concepts are grouped into *p*-value fallacies (**PV.IP** = *inverse probability* fallacy; **PV.R** = *replication* fallacy; **PV.ES** = *effect size* fallacy; **PV.CPS** = *clinical or practical significance* fallacy; **PV.correct** = correct interpretation), confidence intervals (CI), and Bayes factors (BF).

Effects of scale versions. To date, studies investigating *p*-value misconceptions have only surveyed individuals' statements about significant *p*-values (i.e. given $p = .01$ or .001). As we also administered a version of the scale dealing with statements about non-significant outcomes, (i.e. $p = .30$), binary logistic regression analyses (see **Table 2**) were run to compare baseline accuracy rates for both sets of the *p*-value items ("I don't know" responses coded as incorrect; Holm-Bonferroni correction applied for multiple testing). Analyses revealed that individuals were 2.19 times more likely ($p < .001$) to avoid falling prey to the clinical or practical significance fallacy (PV7) when interpreting a false statement about a non-significant outcome (i.e. *"Obtaining a statistically non-significant result implies that the effect detected is unimportant"*) as compared to a significant outcome (i.e. *"Obtaining a statistically significant result implies that the effect detected is important"*). Additionally, inverse probability item PV1 was 1.76 times more likely ($p < .001$) to be correctly recognized as false when in the context of interpreting a non-significant *p*-value (i.e. given $p = .30$ ($\alpha = .05$), *"You have absolutely proven the null hypothesis (that is, you have proven that there is no difference between the population means)"*) than in the context of interpreting a significant *p*-value (i.e. given $p = .001$ ($\alpha = .05$), *"You have absolutely proven your alternative hypothesis (that is, you have proven that there is a difference between the population means)"*).

It should be noted, however, that observed misconception rates were also significantly different for two further items (PV5 and PV8) of which the phrasings were identical in both versions, with the exception of the *p*-value provided in the given information (i.e. *"Let's suppose that a research article indicates a value of p = .001* [vs. $p = .30$] *in the results section (alpha = .05)"*). Specifically, individuals were 1.44 times more likely ($p < .001$) to correctly interpret the *p*-value item about replication (i.e. *"The probability that the results of the given study are replicable is not equal to 1-p."*), and 1.61 times more likely ($p < .001$) to correctly endorse the correct *p*-value

definition (i.e. *"The p-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true"*) when considered in the context of a non-significant outcome. While we did not expect to find differences between scale versions for those *p*-value items which were identically phrased (i.e. PV5 and PV8), such a pattern might be explained by the existence of potential item dependencies, contributing to consistency in item responses: In other words, if for instance the odds of arriving at the correct response are higher in one context (such as for items PV1 and PV7 in the non-significant context), then this might carry over to how other items are solved. On the other hand, observed differences in accuracy between scale versions might also be simply due to random sampling error, i.e. reflective of differences between subsamples despite random assignment (i.e. counterbalancing) across scale versions.

Remaining *p*-value items did not yield significant differences in accuracy between versions. Taken together, there is some evidence to suggest that the context in which a *p*-value is being interpreted (i.e. significant vs. non-significant outcome) may influence the propensity for individuals to fall prey to some common misconceptions, whereby the odds of endorsing a misinterpretation is greater when concerning significant outcomes. However, one should be cautious when speculating about these results as they could reflect spurious differences.

Table 2

Effects of scale version at baseline (*n* = 1,915). Binary logistic regression analyses, comparing effects of scale versions on accuracy rates, i.e. differences in baseline misconception rates given interpretations about a significant (p = .001; reference category) versus non-significant (p = .30) outcomes.

Binary logistic regression estimates								**Model summary**			
Item	**Parameter**	**df**	**Estimate $\hat{\beta}$**	**SE**	**Wald**	***p*-value**	**OR**	**Wald χ^2**	***p*-value**	**$R_N^{2\,a}$**	**$R_{CS}^{2\,b}$**
PV1	Version	1	0.56	0.11	26.36	< .001	1.76	27.44	< .001	.02	.01
	Constant	1	0.02	0.15	0.02	.881	1.02				
PV2	Version	1	0.12	0.10	1.58	.208	1.13	1.59	.208	.00	.00
	Constant	1	0.08	0.14	0.34	.561	1.08				
PV3	Version	1	0.19	0.10	3.89	.049	1.21	3.89	.049	.00	.00
	Constant	1	-0.27	0.14	3.95	.047	0.76				
PV4	Version	1	0.10	0.10	0.96	.326	1.10	0.97	.325	.00	.00
	Constant	1	0.40	0.14	7.99	.005	1.49				

PV5	Version	1	0.36	0.10	13.40	< .001	1.44	13.34	< .001	.01	.01
	Constant	1	-1.00	0.14	49.12	< .001	0.37				
PV6	Version	1	0.05	0.11	0.23	.630	1.05	0.23	.629	.00	.00
	Constant	1	0.91	0.15	34.88	< .001	2.48				
PV7	Version	1	0.78	0.10	59.90	< .001	2.19	65.04	< .001	.04	.03
	Constant	1	-0.86	0.14	37.29	< .001	0.43				
PV8	Version	1	0.48	0.10	21.27	< .001	1.61	21.80	< .001	.02	.01
	Constant	1	-0.11	0.14	0.54	.463	0.90				

[a.] Nagelkerke R-squared. [b.] Cox & Snell R-squared.

Rates of Improvement

Rates of improvement were operationalized as increases in conceptual understanding, entailing both shifts from misconceptions or lack of knowledge (resp. incorrect or "I don't know" responses) to correct interpretations; therefore, for the following analyses, accuracy was computed as proportion of correct responses given all responses (i.e. "I don't know" responses coded as incorrect). *See **online supplementary materials (osf)** for properties of full scale & scale subsets.*

Overall learning. Linear mixed model (LMM) analysis first investigated overall learning effects, across the subset of individuals who completed all six pop quizzes (*n* = 162, after exclusions). LMM with random intercepts was used to regress quiz scores (resp., totals at *pretest*, *post-test1,* and *post-test2*) on *time* (dummy coded categorical predictor with 3 time points (*pre*, *post1*, *post2*); *pre* as reference category). As effects of time on learning might be influenced by individual differences in course duration (i.e. number of days/weeks required to complete course; median course duration = 47.93 days or 6.85 weeks), the model also included an interaction term between *time* and *lag* (where here, lag (continuous, mean-centered) represented total time of completion from *pre* to *post2*, measured in hours). As the model contained three effects of interest, i.e. effect of *time* (at *post1* & *post2*) and interaction effect *time*lag* (see **Table 3**), significance levels were corrected for Type-1 errors using a Bonferroni correction (alpha = .0167). Post-hoc Tukey's HSD analysis was included as a follow-up test, accounting for multiple comparisons (family-wise Type-1 error rate = 5%). LMM analyses (model $R^2_{\text{semi-partial}}$ = .28, ICC = .19) revealed a significant effect of *time*, with improvements in scores significantly increasing from baseline mean score of

8.26 (max. score of 14) to 11.13 at *post1* (mean increase from *pre* to *post1* of 2.87, $p < .001$), and 11.60 at *post2* (mean increase from *post1* to *post2* of 3.35, $p < .001$); improvement from *post1* to *post2* (mean increase of 0.48, $p = .169$) was not statistically significant. Interaction with lag was also non-significant $\hat{\beta} = 1.01e\text{-}04$, $p = .777$).

Table 3 – Overall learning effects (*n* = 162). Improvements in learning for the 14 quiz items, across 3 assessment time points (*pre*, *post1*, *post2*); *pre* as reference category. LMM is summarized for fixed effects parameter estimates, as well as random effects (random intercepts variance ($\hat{\tau}$)).

Fixed effects						
Parameter	**Estimate $\hat{\beta}$**	**SE**	**df**	***T***	***p*-value**	**semi-partial R^2**
(Intercept)	8.26e+00	1.87e-01	339.30	44.28	< .001	
Time (*post1*)	2.87e+00	1.95e-01	319.50	14.75	< .001	.25
Time (*post2*)	3.35e+00	1.95e-01	319.50	17.20	< .001	.20
Time*Lag	1.01e-04	3.58e-04	401.80	0.28	.777	.00
Random effects						
Parameter	$\hat{\tau}$					
ID (Intercept)	2.57					
Residual	3.07					

Quiz-level effects. Next, effects of learning were investigated at the quiz-level, applying once again random intercepts LMM analyses to assess effect of time, and accounting for possible interactions between time and lag (mean-centered). Specifically, eight separate LMMs were run to assess improvements across each of the four subsets of items for first immediate learning (i.e. four LMMs), and then retained learning (i.e. another four LMMs; see **Table 4** for overview); for each set of analyses, Bonferroni Type-1 error correction was applied for 4 analyses with 2 effects of interest each (alpha = .006). Due to response attrition, improvements in learning for concepts covered later in the course are based on smaller *n*s, and reported accordingly.

Immediate learning. Results revealed significant improvements in immediate learning (*pre* to *post1*) across all sets of items: Specifically, main effect of time on quiz scores was most notable for less familiar concepts (i.e. BFs and CIs), respectively subset 3 (n = 478, $\hat{\beta}$ = 1.02, p < .001, $R^2_{\text{semi-partial}}$ = .27) and subset 4 (n = 271, $\hat{\beta}$ = 0.93, p < .001, $R^2_{\text{semi-partial}}$ = .23), and less pronounced for the two *p*-value quizzes, i.e. subset 1 (n = 712, $\hat{\beta}$ = 0.72, p < .001, $R^2_{\text{semi-partial}}$ = .07) and subset 2 (n = 325, $\hat{\beta}$ = 0.38, p < .001, $R^2_{\text{semi-partial}}$ = .07). More concretely, while baseline rates for BF and CI items started off lower (resp. at 0.98 and 1.40 correct, out of max 3.0), scores improved on average by approx. 1 point for both sets of items (i.e. increasing to 2.00 and 2.32 items correct at *post1*, resp.). In contrast, *p*-value scores, which were initially higher at baseline (subset 1 mean score of 3.12 out of 5; subset 2 mean score of 2.34 out of 3) incurred relatively smaller improvements, resulting in scores of 3.85 (subset 1) and 2.72 (subset 2) at *post1*. Across all analyses, interaction between lag and time was non-significant (ps ≥ .564); in other words, the amount of time that elapsed between pretest and the first post-test did not have a statistically significant impact on degree of improvement observed across individuals.

Retained learning. Regarding retained learning (*post1* to *post2*), LMMs demonstrated further positive effects of time on *p*-value items, which were statistically significant for subset 1 (n = 207, $\hat{\beta}$ = 0.38, p < .001, $R^2_{\text{semi-partial}}$ = .03), but non-significant for subset 2 (n = 216, $\hat{\beta}$ = 0.10, p = .007, $R^2_{\text{semi-partial}}$ = .01); both analyses yielding non-significant interaction effects of lag*time (ps > .472). In other words, individuals continued to improve on *p*-value items until *post2*, though learning gains were relatively small. In contrast, for BF items (subset 3, n = 206) and CI items (subset 4, n = 225), neither effects of time nor interaction effect with lag were significant (all ps were larger than the Bonferroni-corrected alpha level of .006), demonstrating retention of learning, i.e. no significant increases or drops in scores at *post2*. Once again, time lag, i.e. the amount of time that elapsed between the two assessments, did not have a statistically significant impact on changes in scores observed across individuals for any of the subsets of items.

Table 4

Quiz-level learning effects. Improvements in immediate and retained learning, across the four subsets of items, corresponding to *p*-values (subsets 1 and 2), Bayes factors (subset 3), and confidence intervals (subset 4). LMMs are summarized: Indices include model R-squared (R_β^2), intraclass correlation (ICC), and random effects, i.e. random intercepts variance ($\hat{\tau}$).

	Immediate learning (*pre to post1)*				**Retained learning (*post1 to post2*)**			
Subset (items)	***n***	R_β^2 [a]	**ICC**	$\hat{\tau}$	***N***	R_β^2 [a]	**ICC**	$\hat{\tau}$
Subset 1 (PV1 – PV5)	712	.07	.37	.82	207	.03	.48	.57
Subset 2 (PV6 – PV8)	325	.07	.07	.08	216	.01	.30	.07
Subset 3 (BF1 – BF3)	478	.27	.00	.11	206	.01	.35	.21
Subset 4 (CI1 – CI3)	271	.23	.00	.13	225	.02	.33	.20

a. R_β^2 = standardized measure of multivariate association between the fixed predictors and the observed outcome (Edwards et al., 2008).

Correlates of Performance

Baseline accuracy rates correlated positively and significantly with self-reported statistics expertise (r = .470, n = 582), as well as level of education completed (r = .198, n = 581). As the course progressed, the relationship between performance and initial expertise rating attenuated (respectively to r = .341 (n = 85) at *post1* and r = .093 (n = 69) at *post2*), whereas correlations with education level remained fairly constant (r = .274 (n = 85) at *post1* and r = .238 (n = 69) at *post2*). Performance (i.e. accuracy levels across all 6 quizzes) was also found to systematically correlate positively with corresponding confidence ratings, respectively: r_1 = .250 (n_1 = 1,915), r_2 = .268 (n_2 = 1,045), r_3 = .082 (n_3 = 621), r_4 = .184 (n_4 = 421), r_5 = .233 (n_5 = 371), and r_6 = .300 (n_6 = 276).

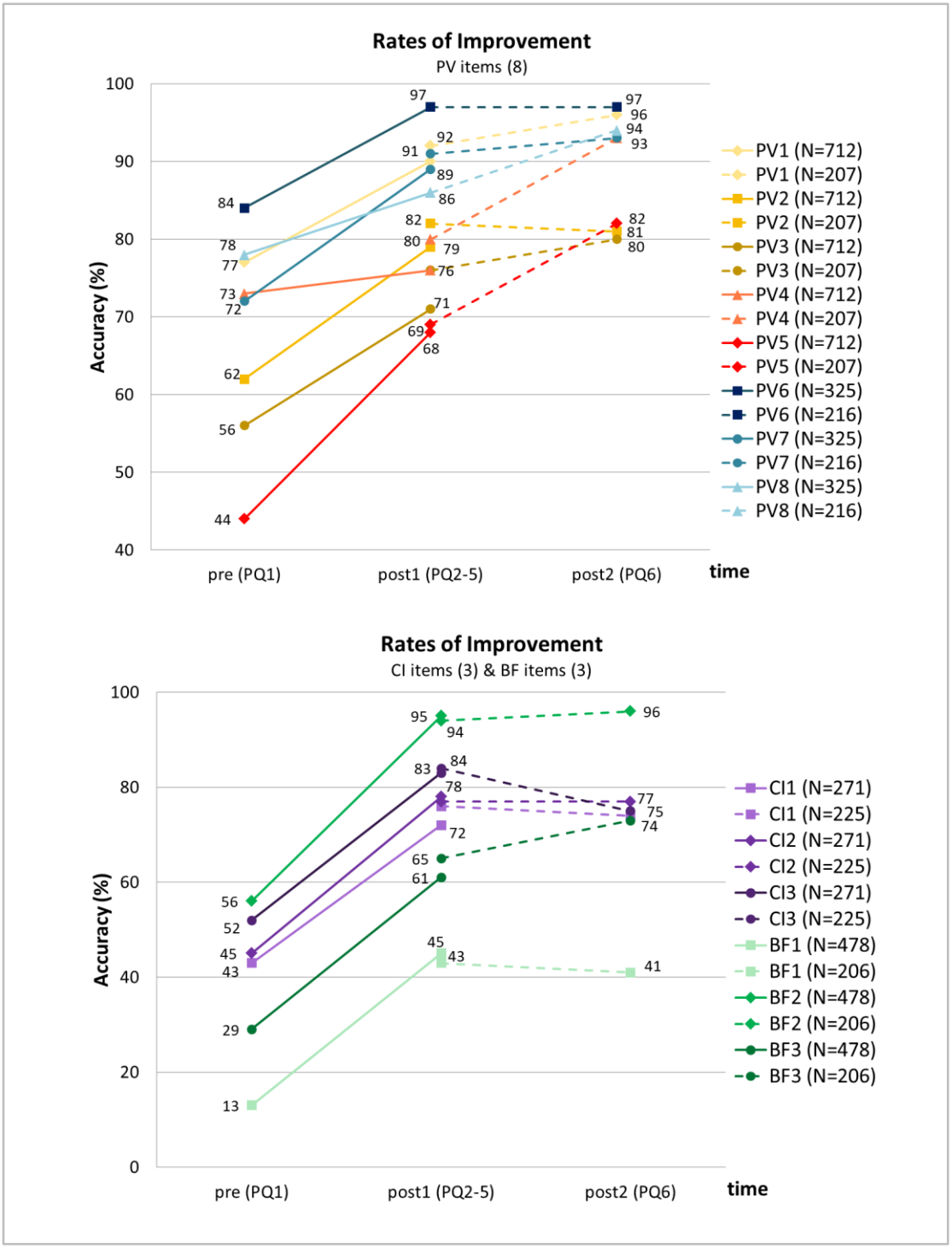

Figure 3 – Rates of improvement (Study 1). Improvements in learning across all 14 items from week 1 (PQ1) to week 8 (PQ6). Graphs demonstrate mean rate fluctuations in immediate learning (solid lines), i.e.from baseline (*pre*) to first post-test (*post1*), and retained learning (dashed lines), from first to second post-test (i.e. *post1* to *post2*). **Top:** 8 *p*-value items (PV1 – PV8), **note:** y-axis ranges from 40% to 100% accuracy; **Bottom:** 3 Bayes factor items (BF1 – BF3) and 3 confidence interval items (CI1 – CI3), **note:** y-axis ranges from 0% to 100% accuracy.

Discussion (Study 1)

Study 1 provided evidence for the ability to improve misconception rates among online learners with respect to interpreting the meaning of *p*-values, BFs, and CIs. Mean scores, across the 14 items, improved from a baseline of 8.26 correct responses to 11.13 (at *post1*) and finally 11.60 (at *post2*); in other words, individuals made on average three to four fewer misconceptions as a result of participating in the 8-week MOOC. Specifically, all *p*-value fallacies (*inverse probability*, *replication*, *clinical or practical significance*, and *effect size*), as well as BF and CI item subsets incurred statistically significant improvements in immediate learning, i.e. as measured immediately after the concept in question was taught, with lesser known items (i.e. BFs and CIs) demonstrating unsurprisingly more pronounced improvements in immediate learning, as compared to *p*-values. Retained learning, as measured by a follow-up assessment in week 8, was observed across all statistical concepts, indicating neither significant increases nor drops in learning from *post1* to *post2*. Taken together, improvements resulting from the MOOC were not only successful in improving learners' ability to correctly interpret statements about statistical concepts, but these improvements were maintained until the final week of the course.

It is worth noting that the MOOC pop quizzes were voluntary (i.e. not graded, nor required to complete in order to advance in or pass the course). As such, it would be of interest to replicate such a study within an educational setting where performance on the quizzes had specific learning goals or consequences. In the same vein, as it is not possible to account for additional inter-individual differences between users (e.g., motivation, interest), or strategies adopted (e.g., working individually vs. in a group, note-taking, etc.), it would be hard to speculate as to whether learning effects reflect conservative versus optimistic estimates with respect to the average learner. On the other hand, as this flexibility in use is integral to the MOOC learning platform, work going forward might seek to investigate which characteristics of a MOOC, that would otherwise not be present in a traditional learning environment, may or may not facilitate the learning process.

One potential limitation of Study 1 was the observation that there may have been too systematic a relationship between the item phrasing and correct answers: Specifically, with the exception of the correct *p*-value interpretation (item PV8), all False items used positive phrasing, whereas all True statements were reverse items. Should participants have noticed this pattern, and used it to drive their responses, this could act as a confound and weaken the assumed benefits of the

MOOC on learning. While we do not expect that this was the case, follow-up work should account for this potential confound. Moreover, one cannot exclude the possibility that mere exposure to items and their correct answers prompted individuals to score higher at subsequent post-tests. That said, were this to have occurred, we should arguably have observed ceiling effects (or at least more pronounced increases in scores, presumably across all items, from *post1* to *post2*), as well as greater improvements for items with a short lag between assessments, neither situation of which was present in our data. Moreover, it should also be mentioned again that individuals were exposed to two scale versions containing non-identical items (i.e. version 1 at *pretest*, version 2 at *post-test1* (order counterbalanced between learners), and finally a combination of items from version 1 and 2 at *post-test2*; see **Appendix A & B**), thus further reducing the likelihood that mere exposure to items wold have realistically accounted for observed improvements. Nonetheless, follow-up work which experimentally manipulates the degree of instructional support learners' receive, might lend explanatory power to the merits of supplementing statistical teaching with explicit tools aimed at tackling and overcoming common misconceptions when drawing inferences from data.

Study 2

As the MOOC was not originally designed to explicitly clarify misconceptions, but rather to teach core concepts, Study 2 investigated the effect of adding teaching material geared toward actively pinpointing and clarifying common *p*-value misconceptions, namely: the *inverse probability*, *replication*, and *clinical or practical significance* fallacies. To this end, in Study 2, we dropped the CI and BF items, and focussed exclusively on the clarification of 9 *p*-value items, comparing improvement rates of controls against an experimental group, who received additional instructional material **(***see* ***online supplementary materials (osf))*** in week 1 of the MOOC.

Procedure

In Study 2, the scale was reduced to the eight original *p*-value items from Study 1 (with some minor modifications), plus one additional item PV9 (*inverse probability*). Slight variations in phrasing, namely for items PV1 (*inverse probability*), PV6 (*effect size*), and PV7 (*clinical or practical significance*), were primarily introduced to create a less systematic relationship between true items and negative phrasing, in response to the concern noted in the Discussion of Study 1. Two

versions of the scale were once again implemented: Version 1 framed interpretations and fallacies in the context of a significant outcome (i.e. $p = .001$; see **Appendix C**), and Version 2 in the context of a non-significant outcome (i.e. $p = .30$; see **Appendix D**). Across all learners, items were presented in a fixed-randomized order, with Version 1 at *pretest* (PQ1; week 1), Version 2 at *post-test1* (PQ2 & PQ3; weeks 1-4), and a combination of both versions at *post-test2* (PQ4; week 8). In order to investigate the effect of explicit clarification on misconception improvements, an experimental group was provided one additional assignment in week 1, as compared to controls, on *"Understanding common misconceptions about p-values"*, which included a series of explanations and practical exercises (e.g., interpreting graphs and use of shinyapps; s*ee* ***supplementary materials (osf)****)*, and which targeted and outlined five common misunderstandings (see below; fallacies where relevant specified in parentheses). Assignment also included 14 multiple choice questions to test students' learning; upon submission, tailored feedback was provided to users in cases where the wrong answer was selected (80% was the required passing rate, multiple response attempts were allowed). Demographics and confidence ratings were again voluntarily requested as in Study 1.

Common misconceptions about *p*-values:

1. A non-significant *p*-value means that the null hypothesis is true. (*inverse probability*)
2. A significant *p*-value means that the null hypothesis is false. (*inverse probability*)
3. A significant *p*-value means that a practically important effect has been discovered. (*clinical or practical significance*)
4. If you have observed a significant finding, the probability that you have made a Type 1 error (a false positive) is 5%.
5. One minus the *p*-value is the probability that the effect will replicate when repeated. (*replication*)

Exclusion Criteria

The same exclusion criteria as in the first study was also applied in Study 2, including response latency outliers (see **Appendix F** for overview of final lag distribution statistics). Additionally, users who did not provide responses for the added assignment in the experimental

condition were excluded from analyses comparing conditions (but included in measures of baseline misconception rates and response trends).

Results (Study 2)

Sample

Total number of learners at *pretest* (*N* = 1,301; *preliminary data*) was split randomly between *c*ontrol (*n* = 652) and experimental (*n* = 649) groups. Study 2 demographics (*n* = 1,195) were quite similar to those of the first study, consisting of 59.3% male participants (38.8% female; 1.9% opting not to specify), with a mean age of 32.90 years (*SD* = 9.72); 31.6% English native speakers, and 80.4% of having previously taken a statistics course before participating in the MOOC. Thirty-six percent of learners rated their statistical expertise-level as beginner, 56.5% as intermediate, and 7.5% as advanced. Academic experience (i.e. degree obtained) was distributed as follows: 7.9% high school diploma, 29.5% bachelor's degree, 43.6% Master's degree, 13.7% PhD degree, 3.6% post-doctoral degree, and 1.7% professorship attained.

Response Trends and Baseline Misconception Rates

Study 2 response trends were similar to those patterns observed in Study 1: High dropout rates were observed from week 1 to week 8, and proportion of correct responses rose (approx. 88-90% correct by week 8) as proportion of incorrect and "I don't know" responses shrunk (see **Table 5**). Visual inspection of **Table 5** also demonstrates some differences between Control and Experimental groups: While at baseline, proportions of correct, incorrect, and "I don't know" responses are fairly matched between conditions, trends across pop quizzes 2 through 4 indicate systematically larger proportions of correct answers among the Experimental respondents, though whether this effect of condition is significant will be more closely assessed via LMM analyses below. In terms of baseline misconception rates, Study 2 findings were also consistent with accuracy rates for *p*-value items observed in Study 1: The *replication* fallacy proved once again to be the most difficult at baseline (51% - 64% accuracy), whereas the *effect size* fallacy was again the least problematic (67% - 85% accuracy; see **Figures 4C & 4D**).

Table 5

Response trends & confidence levels. Accuracy rates (i.e. mean proportion of correct, incorrect, & "I don't know" responses) across all individuals (split between Control and Experimental groups), and mean confidence ratings (CRs), across pop quizzes (PQs) 1 to 4. Items covered in each PQ were clustered into two subsets of *p*-value items: subset 1 (*inverse probability* and *replication* fallacies), and subset 2 (*effect size* and *clinical or practical significance* fallacies, and the correct *p*-value definition).

	PQ1 (*N* = 1,143)		**PQ2** (*N* = 504)		**PQ3** (*N* = 179)		**PQ4** (*N* = 100)	
	Control	**Expt.**	**Control**	**Expt.**	**Control**	**Expt.**	**Control**	**Expt.**
Items covered	Subsets 1 & 2		Subset 1		Subset 2		Subsets 1 & 2	
N	566	577	281	223	90	89	62	38
Correct (%)	59.78	59.92	76.77	89.23	85.57	91.03	87.81	90.07
Incorrect (%)	23.76	24.20	20.40	9.57	12.23	7.87	11.12	8.18
I don't know (%)	16.46	15.84	2.85	1.17	2.20	1.13	1.07	1.74
mean CR *(7-pt Likert)*	4.45	4.48	4.91	5.59	5.24	5.51	5.56	5.76

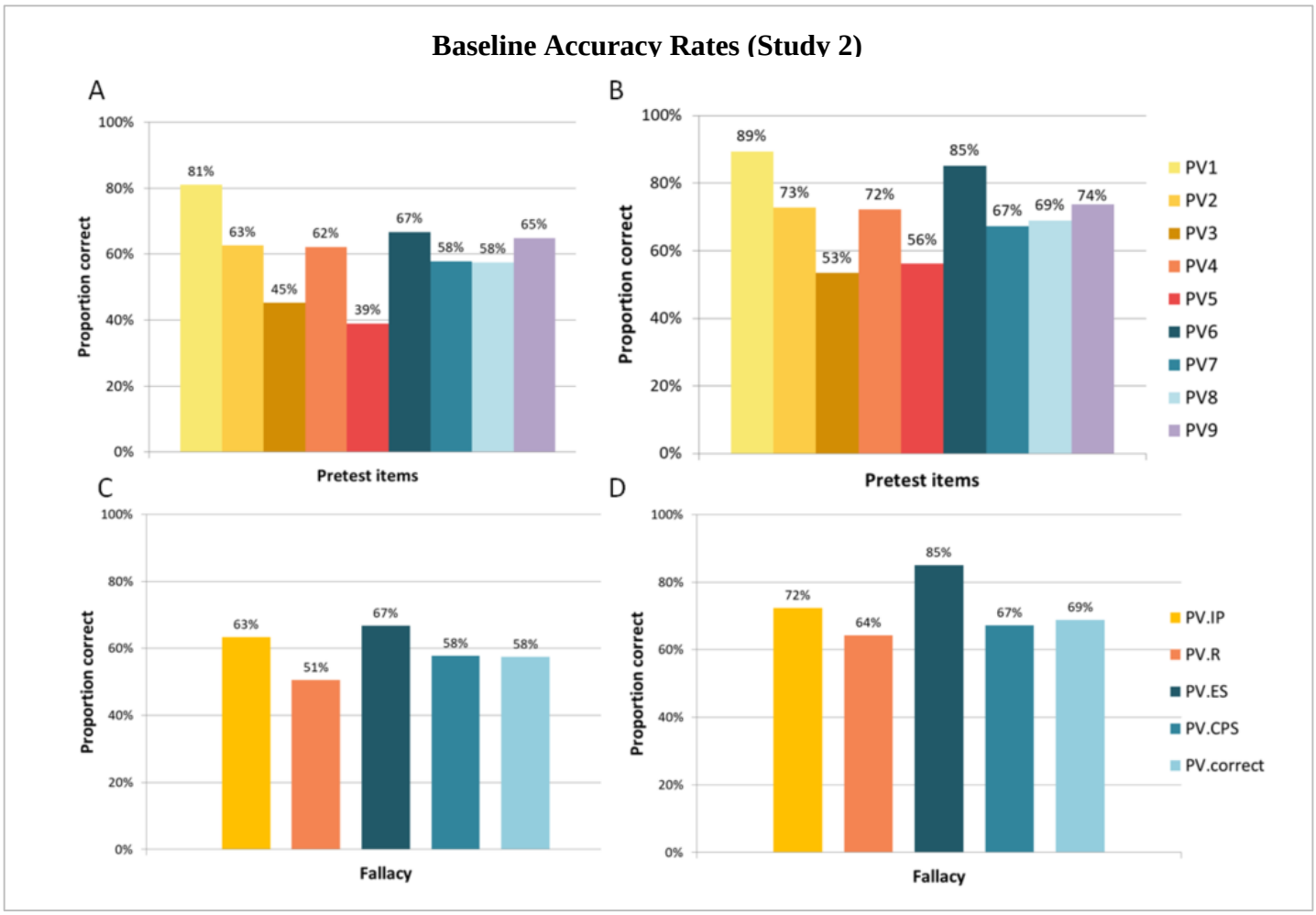

Figure 4 – Baseline accuracy rates (Study 2). Accuracy rates across the 9 individual items (**Figures 4A & 4B**), and averaged across *p*-value fallacies (**Figures 4C & 4D**). Accuracy rates are computed in two manners. **4A & 4C:** As the proportion of correct responses given all provided responses (i.e. "I don't know" responses coded as incorrect); **4B & 4D:** As the proportion of all attempted responses (i.e. "I don't know" responses omitted). Individual items consist of 9 *p*-value items (PV1 to PV9), measuring four distinct *p*-value fallacies (**PV.IP** = *inverse probability* fallacy; **PV.R** = *replication* fallacy; **PV.ES** = *effect size* fallacy; **PV.CPS** = *clinical or practical significance* fallacy; **PV.correct** = correct interpretation).

Rates of Improvement

Overall learning. Linear mixed model (LMM) analysis first investigated overall learning effects, across the subset of 9 *p*-value items, accounting for effect of condition (model R_{β}^{2} = .15, ICC = .32). Across the *n* = 52 individuals who completed all assessments (*pretest*, *post-test1*, and *post-test2*), scores for the control group (n_c = 34) improved from a baseline rate of 6.45correct responses to 7.30 at *post1*, and a final score of 7.81 (out of a max 9.0) at *post2*. Overall, learners in the control condition improved by 1.36 points as a result of participating in the 8-week MOOC ($p < .001$), with those in the experimental group (n_e = 18) displaying a mean further improvement in score of 0.52 points at *post2* (effect of condition positive but non-significant; see **Table 6**). Interactions between time and condition, as well as time and lag, were also non-significant. It should be noted, however, that due to steep response attrition, sample size is only a tiny fraction of the desired sample size (and is – barring the exception of very large effects – expected to be highly underpowered,). It is important to stress here that such a tiny sample size prevents us from drawing meaningful inferences at this juncture; as such, while planned analyses are reported, all preliminary estimates below will be updated after a second wave of data collection and reinterpreted based on a larger and more informative *N*.

Table 6

Overall learning effects (*n* = 52). Improvements in learning for the 9 *p*-value items, across 3 assessment time points (*pre*, *post1*, *post2*); *post2* and control group as reference categories. LMM is summarized for fixed effects parameter estimates, as well as random effects (random intercepts variance ($\hat{\tau}$)).

Fixed effects

Parameter	Estimate $\hat{\beta}$	SE	df	*T*	*p*-value	semi-partial R^2
(Intercept)	7.81e+00	2.79e-01	111.30	27.97	< .001	

Time (*pre*)	-1.36e+00	3.01e-01	99.40	-4.54	< .001	.08
Time (*post1*)	-5.11e-01	3.01e-01	99.40	-1.70	.092	.01
Condition	5.63e-01	4.78e-01	112.70	1.18	.241	.01
Time(*pre*)*Condition	-3.38e-01	5.14e-01	99.72	-0.66	.513	.00
Time(*post1*)*Condition	3.18e-02	5.14e-01	99.72	0.06	.951	.00
Time(*post2*)*Lag	1.90e-04	4.19e-04	126.70	0.45	.651	.00

Random effects

Parameter	$\hat{\tau}$
ID (Intercept)	1.10
Residual	1.53

Quiz-level effects. In order to gain additional insights into learning effects across the two subsets of items, namely with respect to effects of additional training, random intercepts LMMs were once again run to separately assess improvement for immediate learning and retained learning. LMMs included effects of time and condition, as well as interactions between time and lag (mean-centered) and between condition and time; Bonferroni Type-1 error correction was applied for each sets of analyses (i.e. immediate vs. retained, each consisting of two LMMs with four effects of interest per analysis; alpha = .006). Subset 1 consisted of six items measuring the *inverse probability* fallacy (4 items) and *replication* fallacy (2 items), and subset 2 three items measuring the *effect size* fallacy (1 item), *clinical or practical significance* fallacy (1 item), and the correct *p*-value definition (1 item) (see **Figure 5**). Again, it is worth stressing that regarding the analyses below, small sample sizes (especially in the case of retained learning) prevent us from drawing meaningful conclusions at this point; in the case of the immediate learning analyses, *n*'s start to approach a more reasonable size upon which to interpret estimates.

Immediate learning. Results revealed significant improvements in immediate learning (*pre* to *post1*) across both sets of items: Specifically, effect of time on quiz scores was greater for subset 1 ($n = 353$, $\hat{\beta} = 0.64$, $p < .001$, $R^2_{\text{semi-partial}} = .03$), relative to subset 2 ($n = 123$, $\hat{\beta} = 0.37$, $p = .004$, $R^2_{\text{semi-partial}} = .03$), while effect of condition was non-significant for either of the subsets (*p*s > .064). Interaction between condition and time was only significant for improvement in mean scores on the

first set of items targeting the *inverse probability* and *replication* fallacies ($\hat{\beta}$ = 0.71, $p < .001$, $R^2_{semi\text{-}partial}$ = .02), but non-significant for subset 2 targeting *effect size* and *clinical or practical significance* fallacies (p = .848). In other words, while both groups (control vs. experimental) started off with comparable baseline scores for subset 1 (4.05 and 4.13, resp., out of max. 6.0) and subset 2 (2.15 and 2.41, resp., out of max. 3.0), rate of improvement for subset 1 was significantly less pronounced for controls as compared to the experimental group, yielding respective mean scores at *post1* of 4.69 versus 5.48. By contrast, subset 2 displayed overall effects of time, but a non-significant interaction with condition, yielding comparable final mean scores at *post1* of 2.52 and 2.75 (control vs. experimental). Additionally, across both analyses, interaction between lag and time was non-significant (ps > .859): Similar to Study 1, amount of time that elapsed between assessment time points did not have a significant effect on changes in score across either subsets of *p*-value items and for both conditions.

Retained learning. LMMs demonstrated no significant effects of time on retained learning (*post1* to *post2*), in other words no significant increases nor drops in learning occurred across either of the item subsets: subset 1 (n = 56, $\hat{\beta}$ = 0.45, p = .047, $R^2_{semi\text{-}partial}$ = .03), subset 2 (n = 60, $\hat{\beta}$ = 0.07, p = .497, $R^2_{semi\text{-}partial}$ = .00). Effect of condition (ps ≥ .105), as well as interactions between lag and time (ps > .761), and condition and time (ps > .872), were all non-significant. Overall, learning gains held roughly steady until week 8 of the MOOC.

Table 7

Quiz-level learning effects. Improvements in immediate and retained learning, across the two *p*-value item subsets. LMMs are summarized: Indices include model R-squared (R_β^2), intraclass correlation (ICC), and random effects, i.e. random intercepts variance ($\hat{\tau}$).

	Immediate learning (*pre to post1*)				**Retained learning (*post1 to post2*)**			
Subset (items)	***n***	R_β^2 [a]	**ICC**	$\hat{\tau}$	***n***	R_β^2 [a]	**ICC**	$\hat{\tau}$
Subset 1 (PV1 – PV5, & PV9)	353	.14	.17	.55	56	.08	.26	.32
Subset 2 (PV6 – PV8)	123	.07	.09	.07	60	.01	.42	.17

a. R_β^2 = standardized measure of multivariate association between the fixed predictors and the observed outcome (Edwards et al., 2008).

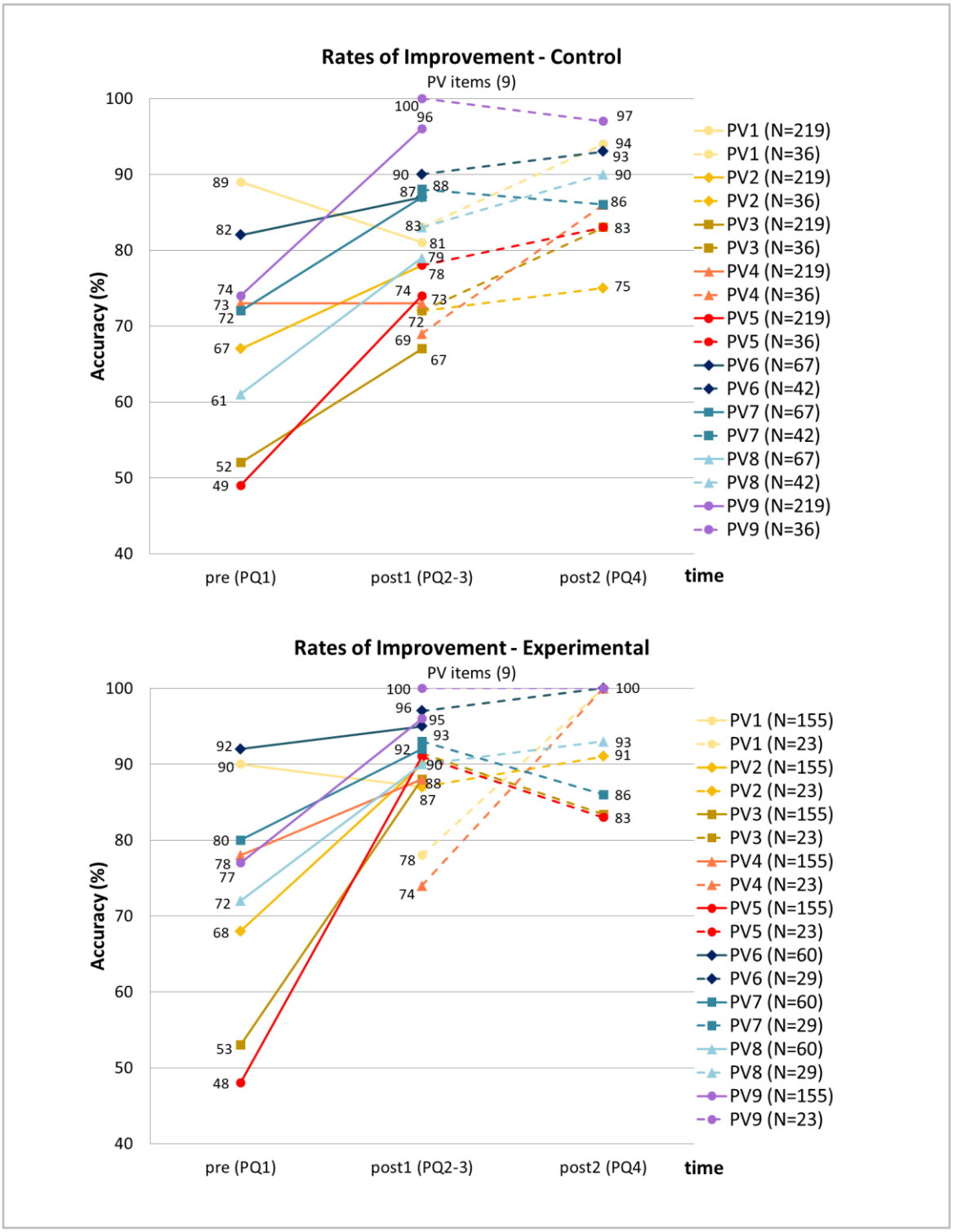

Figure 5 – Rates of improvement (Study 2). Improvements in learning across all 9 items from week 1 to 8. Graphs demonstrate mean rate fluctuations in immediate learning (i.e. from *pre* to *post1*), and retained learning (i.e. from *post1* to *post2*); **note:** y-axis ranges from 40% - 100% accuracy. **Left:** Control condition; **Right:** Experimental condition.

Correlates of Performance

Baseline accuracy rates significantly correlated once again with self-reported statistics expertise (r = .301, n = 1,098), and level of education completed (r = .194, n = 1,098). As in Study 1, performance levels systematically correlated positively with corresponding confidence ratings at each assessment (i.e., r_1 = .283 (n_1 = 1,143), r_2 = .398 (n_2 = 504), r_3 = .207 (n_3 = 179), r_4 = .412 (n_4 = 100). For individuals in the Experimental condition, mean performance at *post 1* (averaged across all 9 *p*-value items) correlated positively with scores on the *p*-value assignment (r = .402, n = 59); in particular, assignment scores were only significantly correlated with accuracy rates for subset 1 (r = .467, n = 186), but not for subset 2 (r = .146, n = 59).

Discussion (Study 2)

Study 2 specifically investigated the effects of supplementary instructional training on rates of learning, when it came to overcoming common misconceptions of *p*-values. Preliminary findings replicated those of Study 1, demonstrating significant improvements in mean quiz scores, across all individuals, from *pre* to *post1*, and subsequent retained learning from *post1* to *post2*, as a result of participation in the 8-week MOOC. While preliminary analyses, assessed across all 3 time points, found no statistically significant effect of condition on overall learning (i.e. improvements in mean scores across the 9 *p*-value items), analyses specifically addressing immediate improvements, across each of the two item subsets, found that effect of additional training was significant for improving the understanding of *inverse probability* and *replication* fallacies (follow-up measure in week 1), but not significant for *effect size* and *clinical or practical significance* fallacies (follow-up measure in week 4). Due to response attrition, and thus small preliminary samples, these results should not be taken as conclusive; with that said, some speculations might be offered about the inconsistency of condition effects across analyses. Perhaps most salient is the fact that the assignment in question provided to the Experimental group addressed five core misinterpretations, two of which were focussed on the *inverse probability* fallacy and one on the *replication* fallacy (i.e. the concepts tested in subset 1), while only one item tackled concepts addressed in subset 2 (namely, the *clinical or practical significance* fallacy). Moreover, the follow-up test for subset 1 occurred immediately after completion of the assignment at the end of week 1, whereas the follow-up test for subset 2 occurred

in week 4 of the course. Finally, as demonstrated in both Study 1 and Study 2, *inverse probability* and *replication* fallacy items tended to be among the more difficult items, allowing for more room for improvement. As such, prior understanding, in combination with both the relevance of the instructional materials to the items in subset 1, as well as the timing of the follow-up assessment, might explain why learners demonstrated greater improvements in learning with respect to specifically the *inverse probability* and *replication* misconceptions as a result of the additional training, and why the effect of condition might have been reduced when learning was averaged across all 9 items (and over a smaller sample of participants). As these results are only preliminary, all effects must be re-assessed at the intended larger sample size.

(General) Discussion

To date, work on statistical misconceptions among academic psychologists has depicted the following consensus: When it comes to statistical indices, such as *p*-values, researchers' are engrained with intrinsic misunderstandings when interpreting data. The current work challenges this point of view, offering empirical evidence that misconceptions are flexible to change, and can even be significantly improved and maintained. Specifically, participants of an 8-week MOOC demonstrated significant improvements in learning across common *p*-value fallacies, BFs, and CIs. Moreover, preliminary evidence suggests that when provided with additional instructional materials, specifically geared toward pinpointing and clarifying common *p*-value misconceptions (e.g., *inverse probability* and *replication* fallacies), learners incurred greater improvements in their quiz scores, in contrast to those who did not receive additional training. In this way, the present work speaks to the merit in developing more dedicated teaching materials, and supplementing the teaching of statistical concepts with targeted training on how *not* to draw conclusions from data, so as to reduce individuals' likelihood of falling prey to common misinterpretations of statistical indices.

An important contribution of Study 1 was the investigation of differences between baseline misconception rates of statistically significant (p=.001) versus non-significant (p=.30) inferences. Of note, individuals made significantly more mistakes when interpreting certain *p*-value items in the context of a significant outcome (vs. non-significant outcome), namely when it came to the *clinical or practical significance* misconception, as well as one of the *inverse probability* items. Such

findings may be attributed to cognitive biases, such as confirmation bias in light of 'significance chasing' (Ware & Munafo, 2015): Due to "the tendency to emphasize and believe experiences which support one's views and to ignore or discredit those which do not" (Mahoney, 1977, p.161), individuals tasked to interpret a significant result may be more prone to prematurely endorsing a conclusion before questioning its limitations than the converse, i.e. when the outcome (e.g., non-significant *p*-value) conflicts with one's expectations. This ties in with the idea of inflated interpretations to which Bakan (1966) alluded: When it comes to interpreting research findings, over-stated generalizations, unwarranted extensions to different inferential levels, and downplaying limitations, can lead to unrealistic conclusions and false endorsements of study outcomes, an abuse of "the realm of qualitative interpretation of quantitative effects" (Ioannidis, 2008, p. 643).

What is worth pointing out in Study 1 was that across the four subsets of items, only the *p*-value items demonstrated further significant improvements from *post1* to *post2*, with subset 1 yielding the relatively larger additional increase in score. This further improvement, especially for *inverse probability* fallacy items, should not be wholly unsurprising given the findings of Kalinowski et al. (2008): Contrasting NHST teaching with the concept of Bayesian inference was found to help students overcome the *inverse probability* misconception. As the MOOC module on Bayesian statistics occurred in week 2, and the concept of equivalence testing (see Lakens, Scheel, Isager, 2018) in week 6 (both occurring between the first and second post-test of the IP items), it is possible that familiarizing learners with these concepts further improved their understanding of what can and cannot be inferred from *p*-values.

Finally, an interesting and unanticipated finding in Study 2, which was in contrast to Study 1, was the decrease in accuracy from *pre* to *post1* for item PV1 (i.e. one of the *inverse probability* items). While all other *inverse probability* items improved between both assessments, the performance on PV1 at *post1*, across both control and experimental groups, dropped relative to baseline performance only in Study 2, which begs the question why this item in question might may have been detrimentally affected relative to the rest of the items measuring the same type of misunderstanding, specifically in the second study alone. While merely speculation, the following explanation might be offered: One key difference in Study 2 was the addition of item PV9, which attempted to provide an umbrella statement on how to (and how not to) interpret *p*-value statements: *"P-values (e.g., p = .001) are statements about the probability of data, not the probability of a*

theory or hypothesis" (version 1). While not addressed explicitly, this item taps into the argument by Falk and Greenbaum (1995), who emphasize "that rejection of H_0 goes along with believing that H_0 is improbable" (p. 81). Specifically, they state: "When a procedure instructs us to reject a hypothesis, in the context of scientific induction, believing that the hypothesis *deserves to be rejected*, namely that it is no longer credible, is inevitable" (p. 81). In this way, it is possible that item PV1 in Study 2 inadvertently became a double-barrelled item if individuals perceived the first half of the item (i.e. *'You have rejected the null hypothesis'*) as implying a statement about the probability of the null (which would be in direct contention with PV9), whereas the second half of the item would be viewed as an accurate description of a significant finding (i.e. *'that is, you have shown that there is a statistically significant difference between the sample means'*). Despite avid recommendations from Falk and Greenbaum, unpacking this misconception, namely the illusion of probabilistic proof by contradiction, is not routine when teaching frequentists statistics. One trick to help learners overcome these implied misunderstandings of NHST might be to equip them with key statements like PV9 (i.e. *p-values are statements about the probability of data, not the probability of a theory or hypothesis*) which demonstrated ceiling effects in understanding. It should be noted however that even though attempts to simplify NHST inferences into accurate and understandable statements might help scaffold learners, too often such short-hands can themselves result in "interpretational overreach and predictable mistakes" (see Spence & Stanley, 2018). Therefore, blanket statements should ultimately serve only as supplemental support when correctly teaching the concept of statistical significance to learners.

Taken together, the current findings emphasize not only the value but the real potential that targeted training techniques, such as online learning platforms, may have on the effective improvement of statistical misunderstandings among researchers and learners. Perhaps more pertinently, our findings also challenge the notion that misconceptions are impervious to correction. While of course a limitation of the work is that it can only speak to individuals' ability to recognize misconceptions at the level of stand-alone theoretical (and often abstract) statements, and not to whether such learning will transfer into practice, it is nevertheless important to first test whether claims about the "persistency and deepness of the misconceptions" (Sotos et al., 2007) are in fact resistant to change when explicitly tackled, and/or simply due to perhaps inadequate instruction on the limitations and common misinterpretations of *p*-values. To date, the tendency to accept that NHST misconceptions are deeply rooted in the minds of scientists appears to be largely a product of

surveys like the Oakes (1986) study, rather than evidence that instructional interventions are ineffective. While there is no doubt that baseline misconception rates continue to be problematic among students and researchers, this does not entail that improvements are impossible; in fact, it calls for increased work on what makes certain instructional interventions (like disrupting the null ritual) effective. All in all, a greater investment in teaching the *correct* use of *p*-values rather than only endorsing alternatives is warranted (Lakens, 2019), whereby pinpointing common misunderstandings may be a critical first step to circumvent the development of misconceptions in the first place.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance.

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*(3), 290-295.

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, *66*(6), 423.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.

Cohen, J. (1990). Things I have learned (so far). *American psychologist*, *45*(12), 1304.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... & Wilson, S. (2007). Statistical reform in psychology: Is anything changing?. *Psychological science*, *18*(3), 230-232.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75-98.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028-17033.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* (Doctoral dissertation).

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*.

George, D., & Mallery, P. (2010). SPSS for Windows step by step. A simple study guide and reference (10. Baskı).

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. *The Sage handbook of quantitative methodology for the social sciences*, 391-408.

Gliner, J. A., Vaske, J. J., & Morgan, G. A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife*, *6*(4), 291-301.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337-350.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1-20.

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse ofp values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037.

Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of evaluation in clinical practice*, *14*(5), 951-957.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759.

Lakens, D. (2019, April 9). The practical alternative to the p-value is the correctly used p-value. https://doi.org/10.31234/osf.io/shm8v

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269.

Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*, *38*(1), 37-45.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764-766.

Lyu, Z., Peng, K., & Hu, C. P. (2018). P-value, Confidence Intervals and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*, 868.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, *1*(2), 161-175.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103-123.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241.

Oakes, M. W. (1986). *Statistical inference*. Epidemiology Resources.

Rinke, E. M., & Schneider, F. M. (2018). Probabilistic misconceptions are pervasive among communication researchers. https://doi.org/10.31235/osf.io/h8zbe

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. *What if there were no significance tests*, 335-391.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61*(4), 293-316.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98-113.

Spence, J. R., & Stanley, D. J. (2018). Concise, Simple, and Not Wrong: In Search of a Short-Hand Interpretation of Statistical Significance. *Frontiers in Psychology*, *9*, 2185.

Vacha-Haase, T., & Ness, C. M. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice, 30*(1), 104.

Ware, J. J., & Munafò, M. R. (2015). Significance chasing in research practice: causes, consequences and possible solutions. *Addiction, 110*(1), 4-8.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129-133.

APPENDIX A
Scale Version 1 (Study 1)

Task: Please read through the following statements, and mark each as True or False. Note that several or none of the statements may be correct. *[Options: True / False / I don't know]*

Given: Let's suppose that a research article indicates a value of $p = .001$ in the results section (alpha = .05).

Subset	Item		Fallacy	Correct A
1	PV1	You have absolutely proven your alternative hypothesis (that is, you have proven that there is a difference between the population means).	Inverse probability	**F**
	PV2	You have found the probability of the null hypothesis being true ($p = .001$).	Inverse probability	**F**
	PV3	The null hypothesis has been shown to be false.	Inverse probability	**F**
	PV4	The *p*-value gives the probability of obtaining a significant result whenever a given experiment is replicated.	Replication	**F**
	PV5	The probability that the results of the given study are replicable is not equal to 1-*p*.	Replication	**T**
2	PV6	The value $p = .001$ does not directly confirm that the effect size was large.	Effect size	**T**
	PV7	Obtaining a statistically significant result implies that the effect detected is important.	Clinical or practical significance	**F**
	PV8	The *p*-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true.	Correct interpretation	**T**
3	BF1	When a Bayesian t-test yields a BF = 0.1, it is ten times more likely that there is no effect than that there is an effect.	N/A	**F**
	BF2	A Bayes Factor that provides strong evidence for the null model does not mean the null hypothesis is true.	N/A	**T**
	BF3	A Bayes Factor close to 1 (inconclusive evidence) means that the effect size is small.	N/A	**F**
4	CI1	The specific 95% confidence interval observed in a study has a 95% chance of containing the true effect size.	N/A	**F**
	CI2	If two 95% confidence intervals around the means overlap, then the difference between the two estimates is necessarily non-significant (alpha = .05).	N/A	**F**
	CI3	An observed 95% confidence interval does not predict that 95% of the estimates from future studies will fall inside the observed interval.	N/A	**T**

APPENDIX B
Scale Version 2 (Study 1)

Task: Please read through the following statements, and mark each as True or False. Note that several or none of the statements may be correct. *[Options: True / False / I don't know]*

Given: Let's suppose that a research article indicates a value of $p = .30$ in the results section (alpha = .05).

Subset	Item		Fallacy	Correct A
1	PV1	You have absolutely proven the null hypothesis (that is, you have proven that there is no difference between the population means).	Inverse probability	**F**
	PV2	You have found the probability of the null hypothesis being true ($p = .30$).	Inverse probability	**F**
	PV3	The alternative hypothesis has been shown to be false.	Inverse probability	**F**
	PV4	The *p*-value gives the probability of obtaining a significant result whenever a given experiment is replicated.	Replication	**F**
	PV5	The probability that the results of the given study are replicable is not equal to 1-*p*.	Replication	**T**
2	PV6	The value $p = .30$ does not directly confirm that the effect size was small.	Effect size	**T**
	PV7	Obtaining a statistically non-significant result implies that the effect detected is unimportant.	Clinical or practical significance	**F**
	PV8	The *p*-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true.	Correct interpretation	**T**
3	BF1	When a Bayesian t-test yields a BF = 0.1, it is ten times more likely that there is no effect than that there is an effect.	N/A	**F**
	BF2	A Bayes Factor that provides strong evidence for the alternative model does not mean the alternative hypothesis is true.	N/A	**T**
	BF3	A Bayes Factor close to 1 (inconclusive evidence) means that the effect size is small.	N/A	**F**
4	CI1	The specific 95% confidence interval observed in a study has a 5% chance of not containing the true effect size.	N/A	**F**
	CI2	To draw the conclusion that the difference between the two estimates is non-significant (alpha = .05), it is necessary that the two 95% confidence intervals around the means do not overlap.[1]	N/A	**F**
	CI3	An observed 95% confidence interval does not predict that 95% of the estimates from future studies will fall inside the observed interval.	N/A	**T**

[1]Note: Typo observed after implementation of the scale. While item correctly signals a misinterpretation about CIs, it is not identical to the misunderstanding measured in Version 1. To correspond, should be corrected to *"To draw the conclusion that the difference between the two estimates is non-significant (alpha = .05), it is necessary that the two 95% confidence intervals around the means overlap."*

APPENDIX C
Scale Version 1 (Study 2)

Task: Please read through the following statements, and mark each as True or False. Note that several or none of the statements may be correct. *[Options: True / False / I don't know]*

Given: Let's suppose that a research article indicates a value of $p = .001$ in the results section (alpha = .05).

Subset	Item		Fallacy	Correct A
1	PV1	You have rejected the null hypothesis (that is, you have shown that there is a statistically significant difference between the sample means).	Inverse probability	**T**
	PV2	You have found the probability of the null hypothesis being true ($p = .001$).	Inverse probability	**F**
	PV3	The null hypothesis has been shown to be false.	Inverse probability	**F**
	PV9[1]	*P*-values (e.g., $p = .001$) are statements about the probability of data, not the probability of a theory or hypothesis.	Inverse probability	**T**
	PV4	The *p*-value gives the probability of obtaining a significant result whenever a given experiment is replicated.	Replication	**F**
	PV5	The probability that the results of the given study are replicable is not equal to 1-*p*.	Replication	**T**
2	PV6	The value for *p* (e.g., $p = .001$) can occur with large as well as with small effect sizes.	Effect size	**T**
	PV7	Obtaining a statistically significant result implies that the effect detected has important and practical impact.	Clinical or practical significance	**F**
	PV8	The *p*-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true.	Correct interpretation	**T**

[1]Note: PV9 was a new item added in Study 2.

APPENDIX D
Scale Version 2 (Study 2)

Task: Please read through the following statements, and mark each as True or False. Note that several or none of the statements may be correct. *[Options: True / False / I don't know]*

Given: Let's suppose that a research article indicates a value of $p = .30$ in the results section (alpha = .05).

Subset	Item		Fallacy	Correct A
1	PV1	You have not rejected the null hypothesis (that is, you have shown that there is a statistically non-significant difference between the sample means).	Inverse probability	**T**
	PV2	You have found the probability of the null hypothesis being true ($p = .30$).	Inverse probability	**F**
	PV3	The alternative hypothesis has been shown to be false.	Inverse probability	**F**
	PV9[1]	*P*-values (e.g., $p = .30$) are statements about the probability of a theory or hypothesis, not the probability of data.	Inverse probability	**F**
	PV4	The *p*-value does not give the probability of obtaining a significant result whenever a given experiment is replicated.	Replication	**T**
	PV5	The probability that the results of the given study are replicable is equal to 1-*p*.	Replication	**F**
2	PV6	A given value for *p* (e.g., $p = .30$) can occur with small as well as large effect sizes.	Effect size	**T**
	PV7	Obtaining a statistically non-significant result implies that the effect detected has no important and practical impact.	Clinical or practical significance	**F**
	PV8	The *p*-value of a statistical test is the probability of the observed result or a more extreme result, assuming the null hypothesis is true.	Correct interpretation	**T**

[1]Note: PV9 was a new item added in Study 2.

APPENDIX E
Response Latencies (Study 1)

Response latencies (Study 1). Lags in time between sets of assessment time points. Latencies are reported for immediate learning, i.e. from *pretest* (*pre*) to first *post-test* (*post1*); for retained learning, i.e. from first *post-test* (*post1*) and second *post-test* (*post2*); and for total time of completion, i.e. from *pre* to second *post2*. Distribution properties (i.e. skewness and kurtosis) and descriptives (i.e. *M, SD, Median, median absolute deviation (MAD), Min, Max*) are provided in minutes, hours, days, and weeks.

Immediate learning								
Latency (***pre* to *post1***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs [a] *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***[b]	***Kurt.***[c]
PQ1 to **PQ2**	712	56.97 *[2.37]*	63.52 *[2.65]*	26.45 *[1.10]*	0.02 *[1.20]*	252.48 *[1.50]*	1.236	0.642
PQ1 to **PQ3**	478	258.00 *[10.75]*	193.01 *[8.04]*	224.19 *[9.34]*	0.02 *[1.20]*	866.72 *[5.16]*	0.857	0.368
PQ1 to **PQ4**	325	545.72 *[22.74]*	324.04 *[13.50]*	531.92 *[22.16]*	0.02 *[1.20]*	1,507.62 *[8.97]*	0.471	-0.077
PQ1 to **PQ5**	271	719.58 *[29.98]*	387.63 *[16.15]*	706.17 *[29.42]*	0.10 *[6.00]*	1,897..22 *[11.29]*	0.291	-0.027
Retained learning								
Latency (***post1* to *post2***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***	***Kurt.***
PQ2 to **PQ6**	207	919.94 *[38.33]*	436.25 *[18.18]*	1,044.95 *[43.54]*	48.95 *[2,937.00]*	1,837.17 *[10.94]*	-0.408	-0.746
PQ3 to **PQ6**	206	758.85 *[31.62]*	379.53 *[15.81]*	882.32 *[36.76]*	1.03 *[61.80]*	1,630.70 *[9.71]*	-0.336	-0.661
PQ4 to **PQ6**	216	500.38 *[20.85]*	296.71 *[12.36]*	530.89 *[22.12]*	0.23 *[13.80]*	1,367.75 *[8.14]*	0.158	-0.324
PQ5 to **PQ6**	225	348.87 *[14.54]*	245.57 *[10.23]*	352.27 *[14.68]*	0.28 *[16.80]*	1,082.88 *[6.45]*	0.520	-0.143
Total time of completion								
Latency (***pre* to *post2***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***	***Kurt.***
PQ1 to **PQ6**	162	1,040.09 *[43.34]*	441.29 *[18.39]*	1,150.23 *[47.93]*	50.78 *[3,046.80]*	2,004.25 *[11.93]*	-0.500	-0.505

a. Median used as measure of central tendency (due to skewed latency distributions).

b. Skewness; range is considered acceptable between -2 and +2 (George & Mallery, 2010).

c. Kurtosis; range is considered acceptable between -2 and +2 (George & Mallery, 2010).

APPENDIX F
Response Latencies (Study 2)

Response latencies (Study 2). Lags in time between sets of assessment time points. Latencies are reported for immediate learning, i.e. from *pretest* (*pre*) to first *post-test* (*post1*); for retained learning, i.e. from first *post-test* (*post1*) and second *post-test* (*post2*); and for total time of completion, i.e. from *pre* to *post2*. Distribution properties (i.e. skewness and kurtosis) and descriptives (i.e. *M, SD, Median, median absolute deviation (MAD), Min, Max*) are provided in minutes, hours, days, and weeks.

Immediate learning								
Latency (***pre* to *post1***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs [a] *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***[b]	***Kurt.***[c]
PQ1 to **PQ2**	374	69.41 *[2.89]*	76.50 *[3.19]*	41.81 *[1.74]*	0.00 *[0.00]*	300.84 *[1.79]*	1.233	0.684
PQ1 to **PQ3**	127	563.53 *[23.48]*	293.69 *[12.24]*	551.08 *[22.96]*	15.97 *[958.20]*	1,360.62 *[8.10]*	0.357	-0.178
Retained learning								
Latency (***post1* to *post2***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***	***Kurt.***
PQ2 to **PQ4**	59	894.64 *[37.28]*	465.32 *[19.39]*	999.79 *[41.66]*	0.14 *[8.40]*	1,995.26 *[11.88]*	-0.074	-0.723
PQ3 to **PQ4**	71	454.78 *[18.95]*	301.11 *[12.55]*	429.17 *[17.88]*	13.61 *[816.60]*	1,277.81 *[7.61]*	-0.518	-0.389
Total time of completion								
Latency (***pre* to *post2***)	***n***	***M*** hrs *[days]*	***SD*** hrs *[days]*	***Med*** hrs *[days]*	***Min*** hrs *[mins]*	***Max*** hrs *[weeks]*	***Skew.***	***Kurt.***
PQ1 to **PQ4**	55	1,021.15 *[42.55]*	474.08 *[19.75]*	1,123.94 *[46.83]*	34.82 *[2,089.20]*	1,998.17 *[11.89]*	-0.159	-0.834

a. Median used as measure of central tendency (due to skewed latency distributions).

b. Skewness; range is considered acceptable between -2 and +2 (George & Mallery, 2010).

c. Kurtosis; range is considered acceptable between -2 and +2 (George & Mallery, 2010).

Project 2

Exploring indices of repeated *k*-fold cross-validation as predictors of replicability

Arianne Herrera-Bennett[1], Chia Wei Ong[1], Moritz Heene[1]

[1]*Department of Psychology, Ludwig Maximilian University of Munich, Germany*

Abstract

The current study is a *Registered Report* (RR) for a large-scale re-analysis of the *Social Sciences Replication Project* (SSRP; Camerer et al., 2018) replication studies, which evaluated the replicability of 21 experiments published in *Nature* and *Science* between 2010 and 2015. Our aim is to first apply a five-repeated 10-fold cross-validation technique to yield indices of model fit and error estimates (i.e. average R-squared, RMSE, MAE), which will then be correlated with index of replication success (i.e. statistical significance indicator). We will also run an exploratory random forest prediction model of replicability using obtained CV indices and observed replicability correlates (i.e. p-values, effect size, and sample size). As such, the overarching goal of the current project is to investigate whether cross-validation indices, as measures of how well the results of a statistical analysis will generalize to independent data, can serve as predictors of study replicability and/or provide corroborating evidence for the link between effect strength, model generalizability, and replicability. The RR submission is exploratory in nature, and consists of a registration prior to re-analysis of the data (osf project: https://osf.io/m4y3w/).

Abstract word count = 180

Keywords: replicability, repeated *k*-fold cross-validation, forest prediction model

Exploring indices of repeated *k*-fold cross-validation as predictors of replicability

Introduction

Replicability and Credibility of Scientific Findings

Replicability of scientific findings is not only considered a *defining feature* or *core principle* of science, but a *crucial component* underlying scientific progress (e.g., Lakatos, 1970; Meehl, 1990; Ioannidis, 2005; McNutt, 2014; Open Science Collaboration (OSC), 2015; Camerer et al., 2018). When it comes to the accumulation of credible scientific evidence, replication serves as an empirical self-correcting mechanism to test and falsify theory (Popper, 1959; LeBel, Berger, Campbell, & Loving, 2017), allowing for the unfolding body of scientific knowledge to align more readily with scientific truth. While methodologically similar (or *direct*) replications can speak to the existence and/or stability of a phenomenon, methodologically dissimilar (or *conceptual*) replications attempt to "investigate the validity and generalizability of psychological phenomena" (LeBel et al., 2017, p. 255). As such, it follows that establishing trust and confidence in observed findings does not hinge single-handedly upon observing whether a replication attempt of some initial finding 'fails' or 'succeeds', but also upon importantly understanding the nature of replicability itself, insofar as replication rates, predictors, and indicators – an area of research which despite marked concern (Ioannidis, 2005; Prinz; Schlange, & Asadullah, 2011; Begley & Ellis, 2012; Pashler & Wagenmakers, 2012; McNutt, 2014) has to date yielded limited evidence (OSC, 2015).

Estimating and Evaluating Replication Success

Efforts to estimate the replicability of scientific findings have risen considerably within the field of psychology in recent years. While such attempts were initially more focussed on trying to formally quantify, or theoretically estimate, the probability that a given study will successfully replicate (e.g., Greenwald, Gonzalez, Harris, & Guthrie, 1996; Sohn, 1998; Posavac, 2002; Macdonald, 2003; Killeen, 2005; Froman & Shneyderman, 2004; Gorroochurn, Hodge, Heiman, Durner, & Greenberg, 2007; Miller & Schwarz, 2011), more recent undertakings have involved collective large-scale efforts to directly and systematically test the extent to which a sample of documented effects within the literature can be successfully replicated in practice. These include the Many Labs 1, 2, and 3 replication projects (Klein et al., 2014, 2018; Ebersole et al., 2016), the *Reproducibility Project: Psychology* (RPP; OSC, 2015), the *Experimental Economics Replication Project* (EERP; Camerer et al., 2016), and most recently the *Social Sciences Replication Project* (SSRP; Camerer et al., 2018).

In spite of such efforts, researchers have yet to arrive at a consensus as to which single measure or set of criteria should serve as a universal standard for evaluating replication success (Gelman & Stern, 2006; Cumming, 2008; Verhagen & Wagenmakers, 2014; OSC, 2015; Simonsohn, 2015; Camerer et al., 2018). Most commonly, replication success has been characterized as a binary index: detecting an effect that is both significant and in the same direction as the original effect, sometimes referred to as the "statistical significance criterion" (Camerer et al., 2018, p. 2). Effect size measures (e.g., relative effect size of replication to original effect) have also been used as complementary continuous measures to assess degree of replicability. Beyond these indices, the RPP, EERP and SSRP projects explored additional measures as potential indicators of replication success, including meta-analytic estimates (combining replication with original effects), use 95% confidence intervals and 95% prediction intervals, small telescopes approach, Bayes factors, Bayesian mixture models, peer beliefs (i.e. prediction markets), and subjective assessments of replicability success by the replication teams (for details, see OSC, 2015 and Camerer et al., 2016, 2018). Overall, no one indicator in any of the projects was found to systematically reflect or sufficiently capture replication success; nor do authors suggest however that the list of potential indicators investigated is exhaustive.

Similarly, when it comes to agreeing upon a shared definition of replication probability, and in turn a formal method to mathematically approximate replication probabilities from experimental data, sharp disagreements continue to persist among researchers (Miller & Schwarz, 2011). Despite nontrivial criticisms (Cumming, 2005; Doros & Geier, 2005; Macdonald, 2005; Wagenmakers & Grünwald, 2006; Iverson, Lee, Zhang, & Wagenmakers, 2009; Iverson, Lee, & Wagenmakers, 2009, 2010; Iverson, Wagenmakers, & Lee, 2010; Maraun & Gabriel, 2010), Peter Killeen's p_{rep} (2005) and $p_{rep,aug}$ (2007) methods are still to date arguably among – if not the most – prominent of approaches to approximate replication probabilities. Moreover, while alternative methods have been proposed (e.g., Iverson, Wagenmakers, & Lee, 2010), lack of general consensus or adequate formal evaluation of such methods, in combination with sheer technical complexity, almost certainly limits the reception and impact of these potential improvements on estimation methods of replicability (Miller & Schwarz, 2011).

Correlates of Replicability: Strength of Initial Evidence

One note-worthy contribution of the RPP, EERP, and SSRP projects was the observation that the strength of the initial finding was found to be more consistently related to, or predictive of, a study's propensity to successfully replicate compared to other indicators, such as characteristics of the research implementation (e.g., quality) or research team (e.g., expertise) (OSC, 2015). Specifically, across all three replication projects (RPP, EERP, SSRP), original *p*-values correlated negatively with replication success (Spearman correlation coefficients: r = -.327, r = -.572, r = -.405, resp.). Moreover, the RPP found replication success to correlate positively with both original effect size (Spearman r = .304), as well as replication effect size (Spearman r = .731), whereas EERP and SSRP reported positive correlations between relative effect size and replicability (Spearman rs = .846 and .842, resp.). This was exemplified in the SSRP results which found that mean relative effect size for the set of effects that replicated was 74.5%, whereas for non-replicating effects it was 0.3%; in other words, "for the non-replicating effects, the mean effect sizes were approximately zero" (Camerer et al., 2018, p. 2). Finally, while the EERP found a positive relationship between original N and tendency to successfully replicate (Spearman r = .627), for the SSRP studies, neither original number of observations, nor number of participants, positively correlated (Spearman rs = -.292 and -.057, resp.). Taken together, results suggest that certain features of the original study, such as strength of initial

evidence (e.g., *p*-value significance and effect size), might serve as important determinants of replicability.

While it goes without saying that this finding affords valuable insight into the nature of replicability, it should not be wholly surprising, especially when considering the nature of effect sizes: "An ES is a measure of the strength of a phenomenon which estimates the magnitude of a relationship" (Kühberger, Fritz, & Scherndl, 2014, p. 1). With this follows the logical assumption that the greater the effect – or the stronger the phenomenon in question – the greater the likelihood that it be detected across a set of investigations. In fact, this assumption is readily verified in the context of meta-analysis, in which "the effect size and number of observed [H0] rejections are positively related" (Francis, 2013, p. 5). In other words, assuming a fixed sample size (and absence of *p*-hacking and publication bias), the stronger the effect under investigation, the greater the proportion of studies in which a significant effect will be detected. Moreover, the concepts of meta-analysis and replicability are inextricably linked, given that a meta-analysis should simply represent the resulting distribution of estimates obtained across all replication attempts, for some given research question or investigated phenomenon.

Taking this one step further, one can appeal to the three-layer model of the research process, a "standard model for the analysis of replication probability" (Miller & Schwarz, 2011, p. 338) used also in meta-analyses (e.g., Wilson & Lipsey, 2001) to quantify sources of variance for a given distribution of observed experimental ESs. The three-layer model identifies three sources of variance, each reflecting one 'layer' or step in the overall research process, namely: the research context, experimental implementation, and data collection steps. Each step introduces, respectively, a source of unique variance, notably the variance in true effect sizes ($\boldsymbol{\sigma^2_T}$), realization variance ($\boldsymbol{\sigma^2_R}$), and sampling error ($\boldsymbol{\sigma^2_E}$). Broadly speaking, the better the approximation of these individual terms, the better the approximation of the resulting overall distribution **O** of observed effects tested (for a detailed overview, see Miller & Schwarz, 2011):

$$\mathbf{O = T + R + E} \; . \tag{1}$$

While we will not go into great detail, there are two points here worth noting. The first is that replicability – or at least the conceptual model of replication probabilities – appears to be, once again, decidedly contingent upon the strength of the true underlying effect. In other words,

"an individual researcher's probability of a positive effect (e.g., across 1,000 replication attempts for the given experiment) depends on the size of the true effect that the researcher selected for study in the first place, so different researchers will have different individual probabilities of observing a positive effect" (Miller & Schwarz, 2011, p. 344).

The second is the concept of 'realization variance', sometimes referred to as 'replication jitter', which is conceptualized as the net perturbation or variation associated with an experiment's implementation across replication attempts (Killeen, 2005, 2007; Miller & Schwarz, 2011). This jitter operates specifically at the level of practical design choices, and can range from trivial or arbitrary differences in experimental setup to intentional methodological deviations from the original design – arguably the layer of variation characterizing where along the continuum from *direct* to *conceptual* a replication attempt lies.

All things considered, given a large true population effect size, there is good reason to expect that it be met with a relatively greater propensity to replicate, than that of a smaller effect. Moreover, theoretically, the stronger the phenomenon in question, the more we would expect it to be more robust against variations in methodological implementation (i.e. replication jitter). With that said, in practice, a myriad of other factors will influence the extent to which conceptual variations in design will yield similar effects; nonetheless, it is important to appreciate these different sources of variance, and how they interact.

Approximating Replication Probabilities

What is interesting to observe is that there exists an asymmetry when it comes to observed correlates of replicability (e.g., *p*-values or ESs) and approximating replicability from that initial data themselves. While strength of original effect appears to be predictive of replication success (i.e. RPP, EERP, SSRP findings), formal attempts to quantify replication probability, as estimated from initial experimental findings, has been often met with the following conclusion: "the data from an initial experiment do not generally provide exact information about the likelihood of a statistically significant replication of that particular experiment" (Miller & Schwarz, 2011, p. 348). In fact, Miller and Schwarz (2011) go so far as to characterize attempts to accurately estimate replication probabilities as "generally unattainable" (p. 337) and "essentially impossible" (p. 355); in sum, stating that "attempting to determine the

individual replication probability associated with a particular new effect seems to be a waste of time, and claiming to have done so is naïve" (p. 357).

The issue lies in the assumption that strength of initial evidence is necessarily an accurate proxy of strength of true effect. As Cumming and Maillardet (2006) astutely point out, even in the (unrealistic) case where the variation between two investigations of the same effect stems only from sampling differences (i.e. population and experimental characteristics are held fixed), the extent to which the original and replication estimates (e.g., means) will agree (or disagree) rests critically on two sources of variance: variation of original estimate around the true mean, and variation of the replication estimate around the true mean. The authors formally define this agreement between estimates in terms of the *capture percentage* (CP), "the percentage of replication means that will fall within a given original CI" (p. 217). CPs, as such, are a direct function of the original estimate, or more specifically how much this estimate deviates from the true population value. Put simply, the extent to which a replication will be deemed successful will depend directly upon how accurately the initial study approximated the population parameter in the first place. Because, however, in real-life, population parameters are unknown, it would be erroneous in practice to take original estimates at face value as being accurate estimates of the true effect against which to judge the replication attempts.

Nevertheless, across a set of studies (as in the case of the RPP, EERP, SSRP), we would still expect to observe a general correlation between studies' effect strength and the propensity to replicate. Despite initial study information being insufficient or inadequate at approximating replication accurately when taken in isolation, general features of a study appear to be more indicative of replicability rates as observed across a distribution of study effects. To this end, we propose to further the investigation of possible determinants of replication success, building directly upon the work of Camerer et al. (2018). Namely, when it comes to study characteristics that might generally forecast the likelihood of replication success, developing or appealing to additional measures of effect strength should contribute to our cumulative knowledge and understanding on how strength of evidence relates to the concept of replicability. As such, we plan to adopt a resampling technique, namely the repeated *k*-fold cross-validation approach, to yield a complementary index of model strength and generalizability for the set of $N = 21$ replication studies in the SSRP data set.

Here it is important to note, for clarity, that while the rationale above for investigating further measures of effect strength did hinge on the idea of *initial* strength of evidence, as well as correlates between *original* study features and replication success, the current project focuses rather on characteristics of the *replication* studies when investigating replicability. As investigations of replicability deal ultimately with the consistency between observed effects, the choice of which estimate (original vs. replication) will constitute the 'initial' evidence is really just a matter of determining the point of reference to judge replication success against. For our study, this initial evidence is captured by the model generalizability indices generated from the replication samples, described in further detail below.

Cross-Validation as Method of Model Generalizability

Cross-validation, which traditionally entails a single split of the data into mutually exclusive training and test sets, involves running the model that was built on the training set on the remaining test set data, and as such is a statistical technique which provides an index of model generalizability. Here, 'generalizability' captures the model's predictive power or validity, i.e. a model's theoretical accuracy or effectiveness in predicting new or unseen data. In other words, cross-validation as a model validation technique should provide an index on how the statistical model will generalize to and perform within an independent data set (for overview of resampling techniques, see e.g., Kohavi, 1995; Beleites et al., 2005; Molinaro, Simon, & Pfeiffer, 2005).

It should then follow logically to bridge the concepts of model generalizability and replicability. That is: The stronger, theoretically, the model prediction accuracy, the smaller the generalization (or out-of-sample) error, and in turn the higher the likelihood that such a model should theoretically hold or fit to an independent sample of data. While sound in rationale, it should be prefaced that while there exists replication jitter from one replication attempt to another, in cross-validation subsets (e.g., training vs. test), this realization variance is necessarily stable across subsamples. This is because ultimately all subsamples are derived from the same overarching sample and subject to the same methodological considerations. As such, variation in population characteristics and practical design decisions (which each introduce their own source of variance), would not be factored into the estimates of model accuracy and generalization error. In fact, it is likely that cross-validation analyses, like bootstrapping, "capitalize during

resampling on the commonalities inherent in a given sample in hand" (Thompson, 1995, p. 92), yielding inflated estimates of model fit and, in turn, inflated estimates of generalizability and/or replicability.

For these reasons, we are particularly interested in also investigating the extent to which some of the theoretical assumptions underlying cross-validation estimates apply in real practice. Specifically, can model validity estimates, which are assessed within a single sample (and are thus reflective of some set of specific sample and design characteristics), in fact generalize to future independent samples? And if so, are such estimates correlated with replication success?

Before outlining our methods, it should made clear that our aim is not to suggest by any means that cross-validation be offered as a substitute to independent replication attempts. In fact, we suspect that generalization of a model beyond the sample of data upon which it was built may be too idealistic a concept when applied in practice. Nevertheless, an index of model strength or generalizability, while perhaps limited, may still provide a general measure of strength of study evidence, and in turn positively correlate with replication success, as in the case of aforementioned observed effect size and *p*-value significance correlates. Therefore, testing this prediction may not only provide corroborating evidence for the link between effect strength and replicability, but also test the strengths or limitations of what can be inferred from cross-validation techniques, namely the five-repeated 10-fold cross-validation approach.

Methods

Brief Study Overview

The current study plans to carry out a set of further analyses on a pre-existing data set, namely borrowing from the *Social Sciences Replication Project* (SSRP; Camerer et al., 2018), which sought to evaluate the replicability of a subset of social science experiments, published in *Nature* and *Science* between 2010 and 2015. The SSRP data set consists of $N = 21$ open source replication studies, with accompanying data sets and analysis scripts for the main analyses investigated (https://osf.io/pfdyw/). In order to evaluate rates of replicability, authors conducted a set of high-powered replication studies, and made use of two primary replication criteria, as well

as a set of six complementary replicability indicators, to assess success of replicated effects as compared to originally reported effects. Primary criterion for replication included a binary measure of success (statistical significance criterion, i.e. detecting a significant effect in the same direction as the original effect, using the same statistical test) as well as a continuous measure of the degree of replication (i.e. relative effect size of the replication).

For our project, we specifically aim to first re-analyze the data applying a five-repeated 10-fold cross-validation technique to yield indices of model fit and error estimates (i.e. average *R*-squared, root-mean-squared error, RMSE, and mean absolute error, MAE) for each of the individual replication data sets. This specific cross-validation approach was selected based on past research showing that, across different resampling techniques (e.g., hold-out, leave-one-out cross-validation (LOOCV), 0.632 bootstrap), repeated 10-fold cross-validation tends to yield the best trade-off between bias and variance (e.g., Breiman & Spector, 1992; Molinaro, Simon, & Pfeiffer, 2005; Kuhn & Johnson, 2013). In other words, both the accuracy of estimation (contingent on training set), as well as the precision of model performance (run on test set), are optimally maximized; i.e. bias (the difference between the average prediction of our model and the actual value) and variance (variance of a model prediction for a given data point) are respectively minimized. This is in part due to the fact that every observation across the full sample contributes with the same weight to both the training and test sets, and thus contributes equally to the error estimation (Beleites et al., 2005). In essence, *k*-fold CV approximates the prediction error as would be obtained via LOOCV, without sacrificing precision for accuracy, and at a much lower computational cost, especially in the case of large samples. Moreover, simulated data has shown that repeated resamplings (in our case, averaging across five repetitions) further reduces bias and variance (Molinaro et al., 2005). Lastly, it should be noted that performance of repeated *k*-fold CV, like most resampling techniques, decreases with lower *n* (see e.g., Kim, 2009). Details of the five-repeated 10-fold CV methodology are included in *Planned Analyses* section below.

Next, these CV indices will be correlated with the binary indicator of replication success as specified in the SSRP study (i.e. statistical significance indicator) as a descriptive analysis. We will then run an exploratory random forest prediction model of replicability using obtained CV indices (i.e. average *R*-squared, RMSE, MAE) and aforementioned replicability correlates

(i.e. p-values, effect size, and sample size): In other words, while the SSRP study investigated replicability predictors independently, we will attempt to investigate how combinations of indicators might serve to jointly predict rates of replicability, assessing alongside the relative weights of their unique contributions using relative variable importance measures (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). As such, the overarching goal of the current project is to investigate whether cross-validation indices, as measures of how well the results of a statistical analysis will generalize to independent data, can serve as predictors of study replicability (i.e. using single-study properties to estimate a study's likelihood to replicate); or perhaps more realistically as corroborating evidence, alongside other indicators (e.g., *p*-values, magnitude of effect), for the link between effect strength and replicability. In this sense, the RR submission is exploratory in nature, and consists of a registration prior to re-analysis of the data (all project information and materials will be made available via osf: https://osf.io/m4y3w/).

Sample

As stated above, the current study borrows from the pre-existing SSRP data set (Camerer et al., 2018), and thus consists of $N = 21$ replication data sets. No additional data collection is involved in the sampling plan. Regarding inclusion and exclusion subject criteria, incomplete or missing data, and/or outliers at the study level, these criteria were already pre-specified by each of the replication teams and integrated into the analysis; as such, the final data sets provided (which will be used for our re-analysis, with no additional edits) have already accounted for necessary exclusions. As each of the $N = 21$ study sets possess unique study characteristics, a breakdown of these individual sample characteristics is provided below (for sample sizes and power, see **Table 1**; for study effect sizes and statistical tests run, see **Table 2**).

It should be noted that the SSRP study ran two stages of replications; this was done to account for potential degrees of inflation of the original reported effect sizes – a factor that the initial *Reproducibility Project* (OSC, 2015) failed to account for. Specifically, Stage 1 replication samples were computed at 90% power to detect an effect 75% the size of the original effect, whereas Stage 2 sample sizes were determined based upon 90% power to detect an effect 50% of the original effect size. In cases where there was failure to replicate at Stage 1, Stage 2 was further administered. This was done to ensure that failure to replicate based on the statistical significance criterion was not due to insufficient power at Stage 1 to detect a smaller effect. On

average, Stage 1 and Stage 2 replication samples were approximately three times and six times larger than original studies, respectively; all in all, five times larger on average. The current study plans to always make use of the larger replication sample sizes when given both options. In other words, in cases where Stage 2 was carried out, we will re-analyze the data on this larger data set; for remaining cases, Stage 1 samples will be use (see **Table 1**, bolded *n* values). The rationale here is so that we can directly compare our results with the final SSRP results.

Table 1

Replication study sample sizes. *Breakdown of N = 21 replication study sample characteristics: original n0, replication samples n1 (Stage 1) and n2 (Stage 2), and accompanying a priori power.*

Studies (N = 21)	Original Study	Replication Study			
		Stage 1		Stage 2	
	n0[a]	*n1*[a]	power[b]	*n2*[a]	power[c]
Ackerman et al. (2010)	54 (54)	259 (259)	.901	**599 (599)**	**.904**
Aviezer et al. (2012)	15 (15)	**14 (14)**	**.930**		
Balafoutas & Sutter (2012)	72 (72)	**243 (243)**	**.898**		
Derex et al. (2013)	51 (366)	**65 (482)**	**.902**		
Duncan et al. (2012)	15 (15)	36 (36)	.909	**92 (92)**	**.906**
Gervais & Norenzayan (2012)	57 (57)	224 (224)	.902	**531 (531)**	**.910**
Gneezy et al. (2014)	178 (178)	**407 (407)**	**.922**		
Hauser et al. (2014)	40 (200)	**22 (110)**	**.919**		
Janssen et al. (2010)	63 (105)	**42 (70)**	**.902**		
Karpicke & Blunt (2011)	40 (40)	**49 (49)**	**.922**		
Kidd & Castano (2013)	86 (86)	285 (285)	.923	**714 (714)**	**.943**
Kovacs et al. (2010)	24 (24)	**95 (95)**	**.923**		

Lee & Schwarz (2010)	40 (40)	123 (123)	.904	**286 (286)**	**.901**
Morewedge et al. (2010)	32 (32)	**89 (89)**	**.904**		
Nishi et al. (2015)	200 (366)	**480 (792)**	**.912**		
Pyc & Rawson (2010)	36 (36)	132 (132)	.904	**306 (306)**	**.901**
Ramirez & Beilock (2011)	20 (20)	26 (52)	.929	**79 (131)**	**.949**
Rand et al. (2012)	343 (343)	1014 (1014)	.920	**2136 (2136)**	**.901**
Shah et al. (2012)	56 (56)	278 (278)	.916	**619 (619)**	**.908**
Sparrow et al. (2011)	69 (69)	104 (104)	.820	**234 (234)**	**.807**[d]
Wilson et al. (2014)	30 (30)	**39 (39)**	**.930**		

Note. Values borrowed directly from Supplementary Tables 3 & 4 (p. 53-54) of the SSRP Supplementary Information.pdf (see https://osf.io/sva2k/).
[a]*n0, n1,* and *n2* = number of observations (number of individuals provided in parentheses).
[b]Stage 1 power = statistical power to detect 75% of the original effect size *r*.
[c]Stage 2 power = statistical power to detect 50% of the original effect size *r*.
[d]Sparrow et al. (2011) is the one case where power is not approx. 90%; power-level was re-computed post-hoc after original authors noted that original *n0* was misreported as *n* = 46, rather than *n* = 69 (see details in Replication Report https://osf.io/84fyw/).

Table 2

Main replication analyses, effect sizes and p-values. *Breakdown of N = 21 replication study effect sizes (with accompanying p-values): original (r), replication at Stage 1 or 2 (r'), and relative replication effect (r'_{rel}). Values are based on respective statistical tests specified.*

Studies (N = 21)	Original Study	Replication Study		
	r [a] (p-value)	r' [b] (p-value)	r'_{rel} [c]	Statistical test
Ackerman et al. (2010)	.270 (4.86e-02)	.063 (1.25e-01)	0.232	Independent-samples t-test
Aviezer et al. (2012)	.961 (3.10e-09)	.829 (1.34e-04)	0.862	Paired-samples t-test

Balafoutas & Sutter (2012)	.278 (1.78e-02)	.146 (2.23e-02)	0.527	χ^2 test
Derex et al. (2013)	.525 (5.41e-05)	.361 (2.96e-03)	0.687	χ^2 test
Duncan et al. (2012)	.674 (4.23e-03)	.436 (1.22e-05)	0.648	Paired-samples t-test
Gervais & Norenzayan (2012)	.289 (2.92e-02)	-.035 (4.15e-01)	-0.123	Independent-samples t-test
Gneezy et al. (2014)	.223 (2.70e-03)	.182 (2.10e-04)	0.818	z-test of proportions
Hauser et al. (2014)	.816 (1.43e-10)	.832 (1.58e-06)	1.020	t-test of regression coefficients
Janssen et al. (2010)	.631 (8.36e-09)	.344 (2.52e-02)	0.545	Mann-Whitney test
Karpicke & Blunt (2011)	.602 (3.93e-05)	.384 (5.89e-03)	0.638	Independent-samples t-test
Kidd & Castano (2013)	.269 (1.33e-02)	-.027 (4.68e-01)	-0.101	2-way btw-subjects ANOVA
Kovacs et al. (2010)	.450 (2.38e-02)	.586 (3.50e-10)	1.301	Paired-samples t-test
Lee & Schwarz (2010)	.388 (1.33e-02)	-.046 (4.36e-01)	-0.120	2-way mixed ANOVA
Morewedge et al. (2010)	.453 (9.21e-03)	.355 (6.49e-04)	0.783	Independent-samples t-test
Nishi et al. (2015)	.201 (4.40e-03)	.116 (1.09e-02)	0.579	t-test of regression coefficient
Pyc & Rawson (2010)	.377 (2.36e-02)	.150 (8.71e-03)	0.398	Independent-samples t-test
Ramirez & Beilock (2011)	.793 (3.02e-05)	-.098 (3.94e-01)	-0.124	2-way mixed ANOVA

Rand et al. (2012)	.141 (8.87e-03)	.026 (2.34e-01)	0.183	z-test of regression coefficient
Shah et al. (2012)	.267 (4.63e-02)	-.015 (7.10e-01)	-0.056	1-way btw-subjects ANOVA
Sparrow et al. (2011)	.368 (1.74e-03)	.050 (4.49e-01)	0.135	Paired-samples t-test
Wilson et al. (2014)	.674 (4.41e-05)	.594 (6.80e-05)	0.880	Independent-samples t-test

Note. Values borrowed directly from Supplementary Tables 3 & 4 (p. 53-54) of the SSRP Supplementary Information.pdf (see https://osf.io/sva2k/), as well D3 – ReplicationResults.csv file (see https://osf.io/abu7k/).
[a] *r* = original standardized effect sizes (*p*-values in parentheses).
[b] *r'* = replication standardized effect sizes after Stage 2; Stage 1 values are reported for those studies that did not proceed to Stage 2 (respective *p*-values in parentheses).
[c] r'_{rel} = relative standardized effect size of replication effect size *r'*; i.e. after Stage 1 and 2 respectively.

Sample Size Rationale

The current project involves analyses **i.** at the individual study level (i.e. repeated 10-fold cross-validation measures), as well as **ii.** analyses across the full set of $N = 21$ studies (i.e., correlations with replicability, and forest prediction model). As such, both levels of analysis are considered with regard to sample size rationale.

Sample size rationale for analyses at individual study level. First, each individual replication study will be re-analyzed in accordance to the main effect under investigation as specified in the original studies (for breakdown of 21 main hypotheses, see Supplementary Table 1, p. 48 of **'SSRP – Supplementary Information.pdf'** https://osf.io/sva2k/). The key difference is that each main analysis will be re-analyzed using a repeated 10-fold cross-validation approach (repeated five times), yielding model fit and error estimate indices (i.e. average *R*-squared, RMSE, and MAE). As such, each study should be sufficiently powered (i.e. at 90%) to run the main analysis that relates to the specific research question of the respective original study. As listed above, the 90% statistical *a priori* power criterion was met in the SSRP replication attempts (see **Table 1**), and pre-specified as a function of each study's main statistical analysis (see **Table 2**), under *Planned Analyses*). It should be noted that there was one exception: in the case of the Sparrow et al. (2011) replication sample, power was re-computed post-hoc at 80.7%

after original authors noted that original *n* was misreported (see Replication Report https://osf.io/84fyw/ for details), and thus does not fit the 90% power criterion. The Sparrow et al. (2011) study will nonetheless be included in the re-analysis; lower power will be considered as a potential constraint in the interpretation of the results.

Moreover, because repeated 10-fold cross-validation requires a sample that can be split into 10 approximately equal subsamples, a large enough initial sample that meets this criteria is needed to carry out the analysis. As such, it is optimal to make use of the $N = 21$ large high-powered samples provided by in the SSRP. It should be noted that some of the sample sizes listed in **Table 1** may look misleadingly small. For example, while the Aviezer et al. (2012) replication sample compares within-subject mean valence ratings (for winning vs. losing expressions) across $n = 14$ subjects, each subject's mean valence rating (for winners vs. losers, resp.) was based on 88 trials (i.e. 176 data points were collected per participant, for a total of 2,464 data points across subjects). This, for instance, differs importantly from the Gervais and Norenzayan (2012) replication sample which compared between-subject mean 'belief-in-God' ratings between two conditions (analytic vs. control). Here, mean group ratings were based on 1 data point per subject, across $n1 = 262$ (analytic group) and $n2 = 269$ (control group) individuals, respectively (i.e. a total of 531 data points across subjects). See breakdown of all 21 design and sample details in **Table A1** of the **Supplementary Materials**.

Sample size rationale for analyses across studies. Analyses across studies will consist of correlations between CV measures and replicability, and random forest prediction model analysis. While an *N* of 21 studies is far from large, it certainly can be considered a valuable starting point considering the constraints that exist when it comes to investigating properties of replicability in practice, and on large-scale high-powered data sets. Of particular merit are indeed the large replication sample sizes which, as the original authors point out, should "get relatively precise estimates of the individual effects of these single replications and the average relative effect sizes" (Camerer et al., 2018, p. 5). As such, we would argue that for both the purposes of the current study, as well as the larger goal of contributing to the accumulating literature on and evidence for potential complementary replicability indicators, the SSRP sample is most apt. Constraints on generalizing obtained results beyond the SSRP sample will be considered when interpreting outcomes.

Experimental Procedure

The experimental procedure for the proposed study concerns first and foremost the steps involved in the preparation of the data, and the re-analysis of the individual replication data sets, before applying the final analyses across studies. Steps include:

1. First, the SSRP database, which constitutes a combination of different data file formats (e.g., .dta, .tsv, .sav, .csv), and analysis syntax formats (e.g., .do, .sps, .R), will be converted into a common format. Namely, all $N = 21$ data sets provided in the SSRP database will be converted (if required) into formats that can be read into R: .csv or .dta for data files, and R code for all analysis scripts. Final set of converted data files and R scripts will be uploaded under the **Materials** component (https://osf.io/ynm8x/) of the main osf project (for an example, see converted Derex et al. (2013) data file https://osf.io/r43zk/ and main analysis syntax file https://osf.io/5mzfs/).

2. In order to ensure that conversion of file formats was successful, we will run the main analyses in both the original file format (e.g., Stata) as well as in the new file format (i.e. R) to check that the original analysis outcomes match (i.e. are reproducible) before proceeding to the re-analysis.

3. Next, the re-analysis component of the project will be specified. First, each main analysis will be converted into an equivalent General Linear Model; re-expressing the analyses as a linear model will allow us to apply the CV technique and estimate the CV indices listed in step 4; this is also required when using the "caret" package (Wing et al., 2018) because it expects regression models as inputs to run the CV analysis. After running a second reproducibility check on the re-expressed analyses (across the full sample), we will specify the five-repeated 10-fold cross-validation analysis in R, using the "caret" package. For each of the 21 individual SSRP studies, the CV analysis will be applied when re-running each respective main replication analysis (see **Table 2** for details of statistical tests). Final set of R scripts to run the CV analyses will be uploaded under the **Planned analyses** component (https://osf.io/6tnk4/) of the main osf project (for an example, see Derex et al. (2013) CV analysis syntax file https://osf.io/m4jfd/).

4. The following indices, resulting from the aforementioned CV re-analyses, will then be computed: average *R*-squared, RMSE, and MAE (details of steps 3 and 4 in *Planned Analyses* section below).

Planned Analyses

The main analyses for the proposed project consist of the **i.** five-repeated 10-fold CV at the individual study level, as well as the analyses across all $N = 21$ studies, including **ii.** correlations with replicability and **iii.** a random forest prediction model of replicability. Each is described in more detail below, specifying where applicable our hypotheses (e.g., **H1**) of the expected results.

Five-repeated 10-fold CV. Briefly, the *k*-fold CV strategy involves randomly splitting the data into k subsets of approximately equal size from which k distinct submodels are built:

> […] *k* different submodels are built by iteratively using $k - 1$ of subsets in each submodel. For each submodel the subset of the data excluded from building the model is used as the test set for that submodel. In this way all samples are used in both model training and testing over the sequence of *k* submodels and the error estimated over the *k* submodels (*k*-fold CV error estimate) provides an estimate of the generalization error of the model built on the entire data set. The test and training sample sizes depend on *k*; common choices are $k = 5$ or 10. In practice, using a single random split of the data is common; however, multiple splits can be done to help control the variance of the estimator. (Beleites et al., 2005, p. 92).

Our paradigm aims to randomly split the data into 10 folds five times over. As such, the generalization error will first be assessed within each of the submodels, before being averaged across all 10 folds. This process will be repeated five times, across which a final set of averaged estimates for all cross-validation indices will be obtained. CV indices in question are: average *R*-squared, RMSE, and MAE. We will also translate this average *R*-squared effect size estimate into its respective standardized effect size (correlation coefficient *r*), which will allow for direct comparison with SSRP original and replication effect size values.

To check for consistency between the cross-validated replication effect size estimates with the initial SSRP replication estimates, we will compute the relative effect sizes of the *r* coefficients obtained via CV (i.e. square-root of average *R*-squared) versus the *r* coefficients estimated from the full sample (i.e. reported in the SSRP output). While we expect to observe some deviations between both sets of estimates, we expect that the mean relative standardized effect size, across all 21 studies, will approximate 1.00, especially given the large replication samples which should yield small standard errors. Such a comparison should serve in a sense as an unbiasedness check regarding the CV effect size averaged estimates in relation to those obtained via the traditional method (i.e. entire sample assessed in one go). Should considerable deviations, however, be observed across any of the individual studies, these differences will be reported.

Correlations with replicability. Replication success across studies will then be correlated with the aforementioned CV indices. Specifically, the replicability indicator is taken directly from the SSRP results, and will consist of the binary statistical significance criterion (i.e. detecting a significant effect in the same direction as the original effect, using the same statistical test), dummy coded (0 = failure, 1 = success). CV indices will be those obtained in the five-repeated 10-fold CV step above, i.e. average *R*-squared, RMSE, and MAE. As these analyses are exploratory, 2-tailed significance tests of the bivariate Spearman correlations (alpha = .05) will be run, consistent with the past investigations of replicability correlates (i.e. RPP, EERP, SSRP). In order to provide a comparison against the initial SSRP findings, correlations will be run across all 21 studies; with that said, assumptions of outliers, normality, linearity, and homoscedasticity will also be checked, and any violations reported. We hypothesize that variance explained (i.e. average *R*-squared) should correlate positively with replicability (**H1a**), whereas fit indices (i.e. RMSE and MAE) should correlate negatively with replication success (**H1b**). Note: While an expected direction is specified for **H1a** and **H1b**, we nonetheless consider such analyses exploratory, hence the aforementioned 2-tailed (alpha = .05) tests.

Random forest prediction model of replicability. Lastly, we will run an exploratory random forest prediction model, using the "caret" package (Wing et al., 2018), applying once again a five-repeated 10-fold cross-validation technique. In other words, across the 21 replication studies, we will generate a classification model, using as the target value the replication success

outcomes (dichotomous index), and as potential predictors (or model 'features') the observed correlates of replicability (i.e. replication study *p*-values, effect sizes, and *N*, where *N* represents the number of observations; see **Table 1**), as well as cross-validation indices (i.e. average *R*-squared, RMSE, MAE). As our model features are continuous, order of features and respective split rules will be determined via a data-driven approach, i.e. by comparing all possible ($n - 1$) single-split partitions as implied by our data, and selecting that which yields the greatest error reduction (i.e. largest information gain), using the information gain classifier criterion. Note: Because this analysis is fully exploratory and data-driven, we do not predict any specific hypotheses regarding the model; all exploratory steps will be reported exhaustively.

Outcome-Neutral Criteria

Regarding potential outcome-neutral conditions (such as absence of floor and ceiling effects), we recognize that given a lack of variation in measures derived from the CV analysis (e.g., floor or ceiling effects for RMSE), this would limit our ability to test the stated hypotheses regarding potential correlates of replicability. With that said, we expect that such an outcome is highly unlikely. In terms of a neutral-outcome, i.e. if we failed to observe any correlations ($rs \sim .00$) between cross-validation measures and replicability index, this would also limit which inferences could be drawn. Namely, lack of evidence for a link between these variables should not be taken as evidence for absence of an effect; rather, results would warrant further investigation before any general claims are drawn. This last point is especially salient given the low number of sampled studies the correlations are being run on.

Preliminary Results[1]

Correlations with Replicability

Preliminary correlation results are reported for $n = 12$ of the SSRP replication studies. Namely, CV indices (i.e. average *R*-squared, RMSE, and MAE) were correlated with replication

[1]As this project is still under Stage 1 review, results reported here are merely preliminary and subject to change post-revision; for the purpose of the thesis, preliminary results are limited to correlations of cross-validation indices and replicability, across $n = 12$ cross-validation re-analyses. The 12 studies reported are arbitrary, i.e. they constitute the subset of studies that were able to re-analyzed in time for the thesis.

success as indexed by the binary statistical significance criterion (i.e. observed replication effect that is both significant ($p < .05$) and in the same direction as the original study). Results are in line with our predictions: *R*-squared correlated positively, whereas RMSE and MAE correlated negatively, with replicability (see **Table 3** for Spearman correlation coefficients). It should be noted that because Spearman correlations were run (pre-registered on account of small sample size and consistent with the correlational approach applied in existing replication studies, i.e. RPP, EERP, and SSRP), resulting coefficients for variables correlated with RMSE and MAE are identical (due to their highly-related nature). In other words, while they do not provide the exact same information (i.e. are not computationally identical), they are nevertheless reflective of the same information, namely the mean residual variance of observed versus predicted data, and were thus highly correlated (Pearson $r = .998$) and in turn provided the same rank information (Spearman $r = 1.000$). In this sense, for our purposes, RMSE and MAE ended up providing redundant information insofar as out-of-sample error indices.

Consistent with existing research, the *p*-value predictor was found to be highly negatively correlated with study replicability. While the RPP, EERP, and SSRP correlated the original studies' *p*-values with replication success, we used the *p*-value of the SSRP replication studies as predictor of replicability (see **Table 3** below). Unsurprisingly, the *p*-value was among the strongest determinants of study replication, alongside *R*-squared; once again, due to the observed high relationship between these two predictors (i.e. *p*-value and *R*-squared), Spearman correlation coefficients for each of these predictors with replication success were equivalent in strength (i.e. $r = .857$), but predictive in opposite directions (i.e. *p*-value as negative predictor; *R*-squared as positive predictor). Taken together, preliminary correlation results agree with the overarching study rationale, namely the observation that the strength of effect appears to be consistently related to a study's propensity to successfully replicate. In sum, the smaller the observed *p*-value, or the greater the variance explained of the predictive linear model (*R*-squared), the greater the likelihood of replication; in contrast, the higher the out-of-sample error of the model (RMSE, MAE), the lower the tendency to replicate.

Table 3

Preliminary correlation results. *Spearman correlation coefficients for n = 21 SSRP replication studies were run between obtained CV indices (average R-squared, RMSE, MAE), replication success (statistical significance criterion), as well as p-values (based on full replication sample).*

	p-value	replication success	*R*-squared	RMSE	MAE
p-value	-	-.857***	-.818**	.210	.210
replication success		-	.857***	-.465	-.465
R-squared			-	-.601*	-.601*
RMSE				-	1
MAE					-

* $p < .05$; ** $p < .01$; *** $p < .001$

Relative Effect Sizes: Replication vs. Cross-validation (CV)

Relative effect sizes were computed to compare estimates obtained via the five-repeated 10-fold cross-validation (CV) approach versus those obtained via the traditional method (i.e. estimated once as based on the entire sample). Specifically, because the CV effect size estimates were necessarily positive values, i.e. calculated by taking the square-root of the *R*-squared value averaged across folds (i.e. sqrt(average *R*-squared)), these were therefore compared against the absolute values of the replication *r* coefficients as reported in the SSRP results (i.e. $| r' |$, where r' corresponds to the replication standardized effect size values as listed in **Table 2** above). Relative effect sizes, across the $n = 12$ studies analyzed, displayed an overall tendency for CV estimates (y-axis) to be approximated as larger than their traditional estimate counterparts (x-axis), whereby only 1 out of the 12 studies yielded a smaller CV estimate, i.e. falling below the diagonal (x = y) line (see **Figure 1**).

While plotted effect sizes in **Figure 1** appear to fall more or less along the diagonal line, the average relative effect size taken across all 12 studies yielded a ratio of 2.16, indicating that on average CV effect size estimates were approximated at roughly twice the size of effect sizes computed via the traditional method. This pattern is not consistent with our predictions which assumed that across all 21 replication studies, deviations in both the positive as well as the negative directions would average out to roughly 0.00, and in turn yield a relative effect size of approximately 1.00. Instead, we observe an almost systematic upward bias across all studies when computing the estimated effect size as the square-root of the CV *R*-squared value, averaged across folds. Certainly we cannot know whether these randomly 12 selected studies arbitrarily resulted in more over- rather than under-estimations of the traditional *r* coefficient, and as such this pattern should be reassessed upon completion of all $N = 21$ study re-analyses.

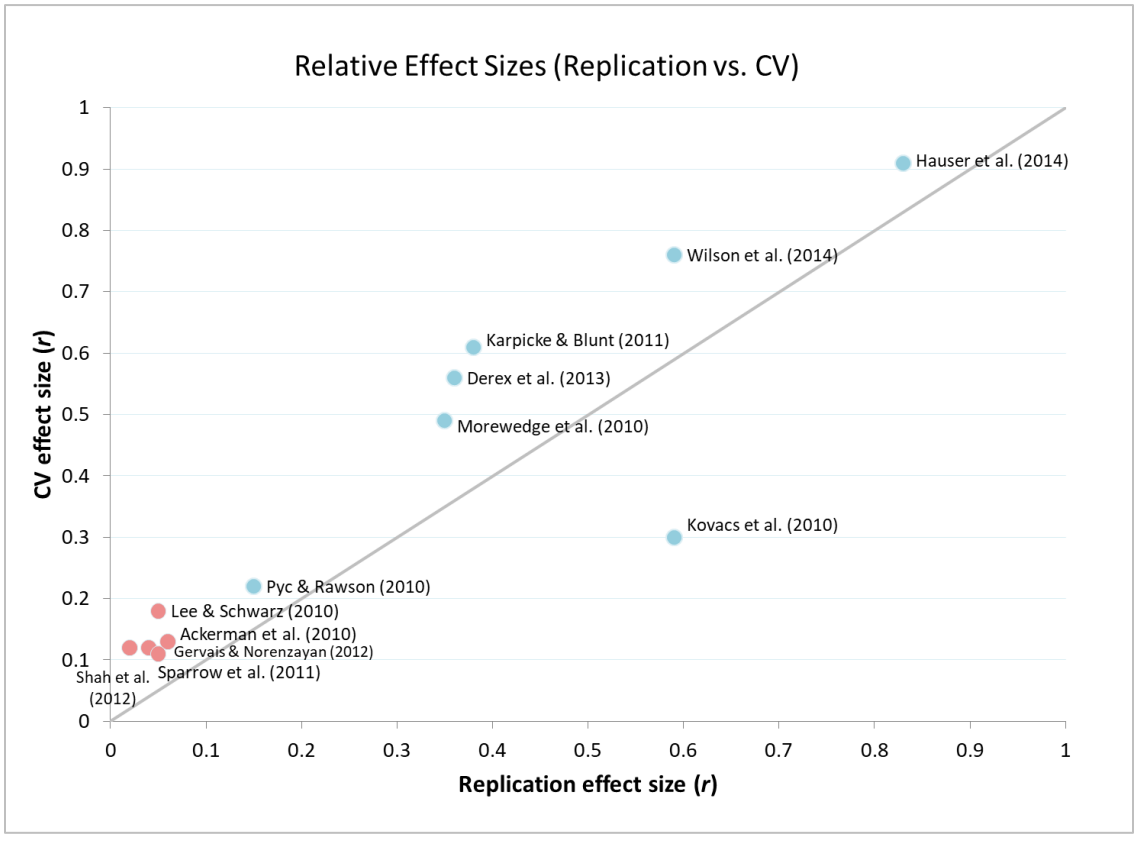

Figure 1 –Replication study effect size versus cross-validation (CV) effect size (correlation coefficient). Absolute values are plotted for replication (i.e. | *r*' |) and CV (i.e. sqrt(average *R*-squared)) effect sizes (x-axis & y-axis, resp.). Diagonal line represents replication effect size equal to cross-validation effect size. Plot is separated by replicated (blue) and non-replicated (red) effects, in accordance with the statistical significance criterion (i.e. $p < .05$ and same direction of original effect).

With that said, there is reason to presume that the resampling process is the source of this upward bias. Most pertinently, across these 12 studies analyzed, re-expressing and re-running the original analyses in the form of linear models on the full sample successfully produced the same effect estimates (e.g., specific t-value), as well as multiple *R*-squared values which (after taking the square-root) were equivalent to the *r* coefficients reported in the SSRP study. Therefore, realistically the reason for these inflated CV effect size estimates is related to what Thompson (1995) referred to about commonalities inherent within subsamples of the full sample, characteristic of resampling; in other words, dependencies unknowingly introduced between training and test sets (Hastie, Tibshirani, & Friedman, 2009; see also Vanwinckelen & Blockeel, 2012). Once all 21 re-analyses are complete, considerations should be made regarding how this non-zero bias may or may not influence the meaning and/or implications of the results.

Discussion

The present study aimed at exploring the relationship between the concepts of model generalizability and study replicability, namely in the context of real large-scale data sets. The rationale for the project stemmed in part from the observation that, across all three recent replication projects (i.e. RPP (2015), EERP (2016), and SSRP (2018)), certain features of the original study findings, indicative of strength of initial evidence, tended to be predictive of study replicability: Namely, statistical significance (*p*-value) correlated negatively, whereas effect size correlated positively, with replication success, whereby the latter was defined by the 'statistical significance criterion', the binary indicator that categorizes a replication as successful if it is significant ($p < .05$) and is in the same direction as the original effect. While not wholly surprising, these findings nevertheless raised the question as to whether further individual study features might serve as important determinants of replicability. For this reason, the current work sought to build directly upon this work and investigate whether estimates of model strength and generalizability might serve as a set of complementary measures to capture strength of evidence and, in turn, predict replication success.

Specifically, we were interested in exploring the use of cross-validation techniques (i.e. five-repeated 10-fold cross-validation) in order to yield indices of model generalizability. The

reason for this boils down as follows: Cross-validation (CV), as a model validation technique, provides an index of how the statistical model will perform within or generalize to an independent data set. While such a generalizability index is established within the two subsets of the same sample (i.e. by assessing how the model developed on the training set performs within the test set), cross-validation is nevertheless assumed to be theoretically predictive of how such a model would hold in unseen data, i.e. beyond that of the training or test sets. In this way, it seemed logical to bridge the concepts of generalizability (obtained via CV) and replication: Namely, the stronger the model prediction accuracy of the pre-registered replication studies (as indexed by average *R*-squared), or the smaller the out-of-sample generalization error (i.e. RMSE and MAE), the higher the expected likelihood of replication. In this way, the current project was also in part motivated by the interest to test this theoretical assumption; in other words, to test in practice the extent to which CV indices (established on the basis of one model, from within one sample) can provide information about how this same model should perform within an independent sample. Thus, by investigating whether CV indices might serve to generally forecast the likelihood of replication success, the current research sought to contribute to a more cumulative understanding of the concept of replicability as well as cross-validation techniques.

In order to accomplish these goals, the present study undertook a re-analysis of the SSRP (Camerer et al., 2018) data set, consisting of $N = 21$ high-powered (90% statistical power) replication studies, whose corresponding original studies were sampled from *Nature* and *Science*, published between 2010 and 2015. Re-analysis across $n = 12$ of the studies (*preliminary results*), in a first step, involved re-expressing each of individual study analyses in the form of a linear model, after which the five-repeated 10-fold CV approach was run to yield three indices of interest per replication study, namely average *R*-squared, RMSE, and MAE (averaged across folds). Preliminary results were consistent with our predictions, indicating high positive correlations (> .80) between replication success and variance explained (i.e. average *R*-squared), and moderate negative correlations (-.47) between prediction error indices (i.e. RMSE and MAE) and replicability. Because RMSE and MAE indices, though not computationally identical, reflect the same information, Spearman rank correlations obtained were identical across these two variables, and as such provided redundant information. While of course these correlation results, taken across the incomplete sample of 12 studies, provide only first insights into our broader investigation, they are nonetheless quite compelling insofar as the observed sizes of the

correlations. Though logically consistent with the theoretical assumptions behind cross-validation, there are nevertheless reasons, namely as it concerns sources of variance (outlined below), which should arguably lead one to be skeptical as to how well model generalizability (as determined from resampling) might translate into practice.

First and foremost, beyond how strong or robust an effect is, the existence of sources of variance when investigating an effect across multiple occasions (barring major methodological deviations or questionable research practices) will necessarily play an important role in how likely a replication attempt will turn out "successful". As exemplified by the work of Cumming and Maillardet (2006), in which the authors identify two important sources of variance that determine the capture percentage (CP) of replication attempts (i.e. variation of original estimate, as well as replication estimate, around the true mean), they highlight the extent to which mere sampling differences influence how much the findings from one replication attempt to the next will agree (or disagree). Consequently, the probability that a replication will succeed (as operationalized by the 95% CI of the replication mean capturing the original sample mean) is a function in part of how accurately the original mean approximated the true mean. Notably, the authors explicitly "assume that replications come from the same population, so variation from sample to sample is caused only by sampling variability, and not by any other changes in population or experimental characteristics" (2006, p. 217). In this way, their simulated replications represent an optimistic set of conditions, not likely reflective of the vast majority (if any) real-life replication experiments. Nevertheless, the idea of CP and sources of variation speak to some important ideas worth considering when it comes to real replication data sets, as in the case of the current project.

Regarding the real SSRP replication data sets, one strong advantage was certainly the large high-powered sample sizes expected to produce "relatively precise estimates of the individual effects of these single replications" (Camerer et al., 2018, p. 5); with that said, the original studies being replicated were not necessarily well-powered, and as such may have produced less precise original estimates of the population parameter in question. What this means in terms of CPs is that by using the less powerful original study to determine the initial mean, which then served as the basis upon which to judge replication success, one may have yielded biased or misleading outcomes if the original estimate strongly deviated from the true

mean (i.e. consequence of variation of original estimate around the true mean). On the other hand, one should not theoretically fare any better given the converse: If the well-powered replication estimate is rather used to establish a presumably more precise initial mean, the CP will nevertheless still be impacted by the potential imprecision of the original estimate, because the CP necessarily functions as a product of both sources of variation.

The reason for stressing this is two-fold: On the one hand, it speaks to a limitation of the current study. Because, as established above, assessing replicability will necessarily depend on both studies being compared, if the original studies are low-powered then this will skew the accuracy of replicability indicators (e.g., whether replication effect falls within the 95% CI of the original effect). While the SSRP project accounted for the shortcoming of the OSC project (i.e. low-powered replication *n*'s), this only took care of one half of the same underlying issue raised. Ultimately, having access to two independent sets of highly powered studies (original vs. replication) would provide for a more valid setup to assess replicability rates of individual studies, and in turn which features of these studies may or may not be predictive of replication success. This is not to say that, under these conditions, the existence of both sources of variance would not continue to play a role, but acknowledging and controlling to some extent these variations in estimates (i.e. aiming for *two* precise estimates) is nevertheless better practice, even if one or both of these effect estimates deviates strongly from the population parameter.

The second reason for raising to the CP concept is because it speaks to a conceptual point which is that in talking about replications, it is commonplace to label the second study as the replication of the first, which is completely logical in the sense that the replication study (as in the case of all three large-scale replication projects) is the outcome of having repeated a previously observed experiment. With that said, when it comes to the statistical concept of replicability, neither the first nor second instantiation of an experiment (let alone a third or a fourth, assuming *all other things equal*) is more representative of the effect in question being approximated. Rather, they each represent one data point along a distribution of effect estimates for the same phenomenon. While perhaps self-evident, this idea of 'strength of evidence' as a predictor or replicability, is thus not constrained to assessing the properties of an original study effect, but rather (as in the case of our project) can also include facets of the replication study. Although our reasons to appeal to the replication data sets were primarily logistical (i.e. requiring

access to the data and syntax, and large *N* suitable for repeated *k*-fold CV), a perhaps more compelling reason to investigate features of both original and replication studies is that they each afford separate (though perhaps related) pieces of information. Future work might consider complementing the assessment of original study properties (to predict replicability of replication studies) with the same approach but to assess the properties of replication studies (to predict replicability of original studies). While results should be consistent, and might appear redundant, they should afford two (rather than one) sets of estimates when attempting to quantify the predictive relationship between some set of individual study properties and replication outcome. It should be noted that here the idea of gaining a more accurate estimate when quantifying predictor estimates is not to say that there is a specific value or cut-off that should determine replicability across studies. Case in point would be the prediction error indices which are unit-dependent (i.e. dependent on the scale the variable of interest is measured on), thus particularly indicative of why specific cut-offs would not be informative across different studies. Rather, this idea would be relevant to meta-scientific projects like the one described below.

One example where this could be implemented is the recent work of Altjmedt et al. (2019) which ran a Random Forest model on 131 original-replication study pairs, combining the results of four large-scale replication projects (i.e. OSC, EERP, Many Labs 1 & 3): Specifically, the authors appealed to over a dozen original study characteristics to predict study replicability (reaching approx. 70% accuracy), observing as top predictors statistical features (i.e. *p*-value, and effect size), but also relatively weaker descriptive variables, such as nature of the finding (interaction vs. main effect), paper length, number of authors, etc. Beyond already providing some interesting insights into the joint predictive power of study features, as well as a first approximation of the weight of each contributing predictor when features are taken in combination, such an investigation might be further exploited by running the same prediction model but on the replication study characteristics, especially given that features of the replication data sets may differ statistically and descriptively from those of the original data sets.

This last point relates in part to another important source of variance which has yet to be discussed, denoted in the introduction as the “realization variance” or “replication jitter”, and conceptualized as the net perturbation or variation associated with an experiment’s implementation across replication attempts. While the SSRP replications were ‘very close’

replications (i.e. highly methodologically similar to the original designs; see LeBel et al., 2017, for replication continuum taxonomy), differences will nevertheless arise from one implementation (of the same design) to the next. Even in cases where these discrepancies could be argued to constitute merely trivial differences, variation of this kind cannot be neglected when it comes to real-life replication research. With regard to the current project, it would be hard to establish the extent to which this realization variance might exist between original and replication attempts, and moreover the degree to which it might affect the performance of CV measures of model generalizability when predicting replicability. It raises the question as to whether bridging the concepts of out-of-sample generalization and replicability makes sense in real practice given more disparate or conceptual replication attempts. Thus while we might be compelled to look positively upon the moderately- to large-sized correlations observed (approx. .47 – .86), these may reflect optimistic estimates yielded under more favourable or unrealistic circumstances. With that said, one interesting and perhaps encouraging finding that Altjmedt and colleagues (2019) pointed out was the fact that "some features that vary across studies are *not* robustly associated with poor replication: These include measures of language, location and subject type differences between replication and original experiments, as well as most of the variation in compensation" (p. 11). In other words, despite discernible differences between studies, at the level of practical design choices and study implementation, these discrepancies were not observed to be predictive low replicability. Though once again these results were based on replication studies where replication teams contacted original authors for original materials and consultation on methodological deviations that were often endorsed. As such, there is reason to reserve skepticism regarding the nature of our preliminary results insofar as how they might generalize beyond the specific research context.

Further Study Limitations & Future Directions

Beyond those already discussed above, another limitation of our preliminary results is that they currently only speak to the predictive power of study features assessed separately rather than in combination. Therefore, running the planned forest prediction model with multiple predictors may provide a more informative overview of which study characteristics are predictive of replication. Additionally, the present study made use of only a single replication indicator (i.e. the statistical significance criterion). As there is no universal standard for

evaluating replication success, and given existing alternatives (e.g., replication mean contained within 95% CI of original mean), our plan going forward will be to make use of complementary indicators (binary and continuous) to assess replicability, when completing and updating the planned correlation analyses as well as when running the forest prediction models. Regarding the RMSE and MAE predictors, aside from providing essentially redundant information, these measures should also be acknowledged as being unit-dependent which itself may skew the results, given that the variables across the set of replication studies is not consistent. As such, another planned deviation from the original design going forward will be to additionally include a standardized version of these measures (e.g., dividing the RMSE by the mean of the outcome variable, such that it can be interpreted rather in terms of the percentage of the mean). Finally, a last point of consideration concerns the observed relative effect sizes, which indicated systematically larger estimates when computed via the repeated *k*-fold CV approach (i.e. multiple estimates averaged across subsample folds) as compared to the traditional method (i.e. single estimate calculated from the full sample). Delving deeper into possible explanations for this upward bias, and its implications regarding cross-validation techniques, would be of further interest.

References

Altmejd, A., Almenberg, A. D., Forsell, E., Ho, T. H., Huber, J., Imai, T., ... & Camerer, C. (2019). Predicting the Replicability of Social Science Lab Experiments.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531.

Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M. G. (2005). Variance reduction in estimating classification error using sparse datasets. *Chemometrics and intelligent laboratory systems*, *79*(1-2), 91-100.

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International statistical review/revue internationale de Statistique*, 291-319.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*(12), 1002-1004.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall?. *Psychological methods*, *11*(3), 217.

Doros, G., & Geier, A. B. (2005). Probability of replication revisited. *PSYCHOLOGICAL SCIENCE-CAMBRIDGE-*, *16*(12), 1005.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*(5), 153-169.

Froman, T., & Shneyderman, A. (2004). Replicability reconsidered: An excessive range of possibilities. *Understanding Statistics*, *3*(4), 365-373.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331.

Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M., & Greenberg, D. A. (2007). Non-replication of association studies:"pseudo-failures" to replicate?. *Genetics in Medicine*, *9*(6), 325.

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated?. *Psychophysiology*, *33*(2), 175-183.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Iverson, G. J., Lee, M. D., & Wagenmakers, E. J. (2009). p rep misestimates the probability of replication. *Psychonomic Bulletin & Review*, *16*(2), 424-429.

Iverson, G. J., Lee, M. D., & Wagenmakers, E. J. (2010). The random effects prep continues to mispredict the probability of replication. *Psychonomic bulletin & review*, *17*(2), 270-272.

Iverson, G. J., Lee, M. D., Zhang, S., & Wagenmakers, E. J. (2009). Prep: an agony in five fits. *Journal of Mathematical Psychology*, *53*(4), 195-202.

Iverson, G. J., Wagenmakers, E. J., & Lee, M. D. (2010). A model-averaging approach to replication: The case of prep. *Psychological Methods*, *15*(2), 172.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological science*, *16*(5), 345-353.

Killeen, P. R. (2007). Replication statistics. *Best practices in quantitative methods*, 103-124.

Kim, J. H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, *53*(11), 3735-3745.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology*.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS one*, *9*(9), e105825.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*(Vol. 26). New York: Springer.

Lakatos, I. (1970). Falsification and the methodology of scientific research. IN LAKATOS, I. & MUSGRAVE, A.(Eds.) Criticism and the Growth of Knowledge.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional.

Macdonald, R. R. (2003). On determining replication probabilities: Comments on Posavac (2002). *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1), 69-70.

Macdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions. *PSYCHOLOGICAL SCIENCE-CAMBRIDGE-*, *16*(12), 1007.

Maraun, M., & Gabriel, S. (2010). Killeen's (2005) prep coefficient: Logical and mathematical problems. *Psychological methods*, *15*(2), 182.

McNutt, M. (2014). Reproducibility.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108-141.

Miller, J., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological methods*, *16*(3), 337.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301-3307.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530.

Popper, K. R. (1959). *The logic of scientific discovery*. University Press.

Posavac, E. J. (2002). Using p values to estimate the probability of a statistically significant replication. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(2), 101-112.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets?. *Nature reviews Drug discovery*, *10*(9), 712.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, *26*(5), 559-569.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8*(3), 291-311.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, *55*(1), 84-94.

Vanwinckelen, G., & Blockeel, H. (2012, May). On estimating model accuracy with repeated cross-validation. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning* (pp. 39-44).

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457.

Wagenmakers, E. J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*(7), 641.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological methods*, *6*(4), 413.

Wing, M. K. C. from J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., … Hunt, T. (2018). caret: Classification and Regression Training. Retrieved from https://CRAN.R-project.org/package=caret

Supplementary Materials

Table A1

Replication studies' sample & design details. *Breakdown of N = 21 replication study designs, namely as they pertain to overall sample sizes (N_p = number of participants), group sizes across conditions (e.g., n1 vs. n2), and number of trials.*

Studies (N = 21)	Replication Study
	Sample / design details
Ackerman et al. (2010)	N_p = 599 participants (between-subject design): 1 composite score per subject (across 6 different item ratings), comparing between *heavy* group (*n1* = 296) vs. *light* group (*n2* = 303) scores.
Aviezer et al. (2012)	N_p = 14 participants (within-subject design): 176 trials (ratings) per subject, comparing mean ratings for *winner* trials (*n* = 88) vs. *loser* trials (*n* = 88).
Balafoutas & Sutter (2012)	N_p = 243 participants (between-subject design): 1 choice made per subject, comparing *control* (*n1* = 120) vs. *preferential* (*n2* = 123) choice frequencies.
Derex et al. (2013)	N_p = 482 participants (between-subject design), split into 4 group sizes: groups of 2 (17 grps), 4, 8, and 16 (16 grps each). Group probabilities of engaging in both tasks (among last 3 trials) are compared.
Duncan et al. (2012)	N_p = 92 participants (within-subject design): 1,276 total trials per subject, presented as follows: 676 are initially *new* object trials, of which 76 are not shown again, 200 are repeated (i.e. *old* trials), and 400 are presented again in a manipulated form (i.e. *similar* trials). Comparison is made between the fractions of objects rated as *similar* when preceded by *new* vs. *old* trials.
Gervais & Norenzayan (2012)	N_p = 531 participants (between-subject design): 1 trial (self-rating) per subject, comparing *'belief-in-God'* scores between *analytic* group (*n1* = 262) and *control* group (*n2* = 269).

Gneezy et al. (2014)	N_p = 407 participants (between-subject design): 1 choice made per subject, comparing percentage choosing to donate to *"charity: water"* in *"50% overhead"* (*n1* = 205) vs. *"50% overhead, covered"* (*n2* = 202) conditions.
Hauser et al. (2014)	N_p = 110 participants (between-subject design), split into groups of 5: 1 score per group, comparing nb of groups 'sustained' in the *unregulated* (11 grps; *n1* = 55) vs. *voting* (11 grps; *n2* = 55) conditions.
Janssen et al. (2010)	N_p = 70 participants (between-subject design), split into groups of 5: 3 observations per group; comparison of average group net earnings between *NCP-C* (11 grps) vs. *C-NCP* (3 grps) conditions.
Karpicke & Blunt (2011)	N_p = 49 participants (between-subject design): Comparison between mean retrieval scores for *retrieval practice* (*n1* = 23) vs. *concept-mapping* (*n2* = 26) groups.
Kidd & Castano (2013)	N_p = 714 participants (between-subject design): Comparison between mean RMET scores (across 36 trials) for *fiction* (*n1* = 349) vs. *non-fiction* (*n2* = 365) groups.
Kovacs et al. (2010)	N_p = 95 participants (within-subject design): 20 trials per subject, comparing mean RTs for the *"P-A-"* trials (*n* = 5) vs. *"P-A+"* trials (*n* = 5).
Lee & Schwarz (2010)	N_p = 286 participants (between-subject design): Comparison of rank difference between chosen and rejected CDs, before vs. after hand-washing activity, for *hand-washing* (*n1* = 147) vs. *no hand-washing* (*n2* = 139) conditions.
Morewedge et al. (2010)	N_p = 89 participants (between-subject design): Comparison between average grams of M&Ms consumed in *control* (*n1* = 44) vs. *30-repetition* (*n2* = 45) conditions.
Nishi et al. (2015)	N_p = 792 participants (between-subject design), split into group sessions (approx. 16.5 subjects per session): Comparison of mean Gini coefficient per session (averaged across 10 rounds per session) for *visible* (*n1* = 24 sessions) vs. *invisible* (*n2* = 24 sessions) treatment conditions.

Pyc & Rawson (2010)	N_p = 306 participants (between-subject design): Comparing mean fraction of recalled mediators (for 48 translation pairs) between *"test-restudy"* (*n1* = 156) vs. *"restudy"* (*n2* = 150) conditions.
Ramirez & Beilock (2011)	N_p = 131 participants (between-subject design), split between *high-pressure (N1* = 79) and *low-pressure* (*N2* = 52) groups. Comparison between pre-/post-test (40 items each) improvement in math performance for *expressive* (*n1* = 34) vs. *control* (*n2* = 45) conditions within the *high-pressure* group.
Rand et al. (2012)	N_p = 2136 participants (between-subject design): Comparison of mean donation in public goods game between *"promote intuition"* (*n1* = 1058) vs. *"promote reflection"* (*n2* = 1078) conditions.
Shah et al. (2012)	N_p = 619 participants (between-subject design), 80 trials per subject: Comparison between mean performance on Dots-Mixed task for *poor* (*n1* = 298) vs. *rich* (*n2* = 321) conditions.
Sparrow et al. (2011)	N_p = 234 participants (within-subject design), 48 trials per subject: Comparing mean color-naming RTs in Modified Stroop Task for *computer* words (*n* = 16 trials) vs. *unrelated* words (*n* = 32 trials).
Wilson et al. (2014)	N_p = 39 participants (between-subject design): Comparing average self-rated enjoyment scores (mean across 3 items) for *"external activities"* (*n1* = 20) vs. *"standard thought"* (*n2* = 19) conditions.

Note. Details borrowed from each of the 21 respective Replication Reports (Post-Replication) (see https://osf.io/sbru6/).

GENERAL DISCUSSION

The present dissertation was broadly concerned with two core questions: First, *how* do researchers evaluate statistical evidence when drawing inferences from data? And how can we *improve* the process of statistical inference-making among researchers in psychology? **Project 1** took a more literal approach to these questions, assessing the effectiveness of an online learning platform in improving baseline misconception rates of statistical indices among learners, as well as measuring whether immediate learning was subsequently retained across the 8-week timeframe. **Project 2**, in comparison, appealed to a meta-scientific design, in order to explore how the use of advanced statistical resampling techniques might be used to gain more informational value about features of individual study effects, and in turn be used as a novel method to predict replicability in large-scale real replication data sets. Both projects and their main results are briefly summarized and discussed below, including implications for instruction, as well as a means to conceptually and practically bridge the two projects together.

5.1. Project 1: Discussion

Project 1 of the dissertation, **"Improving statistical inferences: Can a MOOC reduce statistical misconceptions?"** (Herrera-Bennett, Lakens, Heene, & Ufer), which consisted of two studies, used a repeated-measures design (across three time points) to assess baseline misconceptions rates, as well as improvement rates, of three statistical concepts (i.e. *p*-values, confidence intervals (CIs), and Bayes factors (BFs)), in the context of an 8-week massive open online course (MOOC). **Study 1** challenged an assumption commonly expressed within the literature in response to the observed widespread misuse of null hypothesis significance testing

(NHST) among researchers, that is: that statistical misconceptions are impervious to change. Not only did this first study show that misconceptions did *not* withstand instruction, demonstrating significant improvements in immediate learning across all three concepts, but that gains in learning were able to be retained until the end of the 8-week course. **Study 1** also extended the past research in a few key ways: First, beyond assessing misconceptions across a range of time, rather than only at one single time point, it also provided preliminary evidence of distinctions between the learning trajectories of *p*-values, CIs and BFs: While, unsurprisingly, the CI and BF concepts proved to be the least familiar across the majority of learners at baseline, they incurred on average the greatest overall improvements in immediate learning, with learning being maintained (i.e. non-significantly increasing or dropping) until week 8. In comparison, immediate improvements rates across all *p*-value fallacies were also significant, but relatively less steep on average than CI and BF rates. Interestingly, however, *p*-value items that measured the *inverse probability* and *replication* fallacies, continued to significantly improve until week 8 (though this increase was small). This last finding could be interpreted as corroborating the findings of Kalinowski et al. (2008), and in line with the idea of 'insight by comparison': Because the concepts of Bayesian inference and equivalence testing, introduced between the first and second post-test, are two concepts which offer additional ways to understand how or why a *non-significant* effect is not the same thing as a *non-existent* effect, it is possible that these added conceptualizations may have elicited a deeper understanding in some individuals as to why the *inverse probability* fallacy is in fact incorrect; in other words, why the probability of the data cannot be used to infer the probability of a theory or hypothesis.

Another novel contribution of **Study 1** was the comparison of baseline misconception rates for *p*-value fallacies when considered in either the context of a significant versus non-

significant outcome; in other words, comparing whether individuals were more or less prone to fall prey to certain fallacies when the *p*-value being interpreted was statistically significant (p = .001) versus non-significant (p = .30). Results showed that in some cases, interpretations made in the context of a non-significant outcome yielded fewer misconceptions, whereas the converse was never observed. The most pronounced case was for the *clinical or practical significance* fallacy, which demonstrated that individuals were over twice more likely to correctly recognize the following item as false *"Obtaining a statistically non-significant result implies that the effect detected is unimportant"*, as compared to its counterpart *"Obtaining a statistically significant result implies that the effect detected is important"*. Findings were speculated to support the idea of cognitive biases (e.g., confirmation bias) acting on the process of drawing inferences: If considered in the context of a competitive academic environment, it is plausible that when faced with results that align with desired expectations (e.g., significant effects), researchers might be less inclined to challenge statements that colour results in a favourable light; conversely, if results conflict with expectations or are perceived as undesirable (e.g., non-significant results), these cases may trigger researchers to question the validity of unfavourable statements, or perhaps make them more prone to rationalize why they may not be true.

As the MOOC was not originally designed to explicitly clarify misconceptions, but rather to teach core concepts, the second study of the project investigated the effect of explicit training on learning, focussing exclusively on the improvements in *p*-value items. **Study 2** adopted once again a repeated-measures design but with the addition of an experimental group who received supplementary instructional support in week 1 of the MOOC, namely an extra assignment which explicitly pinpointed and explained common *p*-value misconceptions, and provided active training on how to recognize and avoid them. While results were only preliminary, some support

was found for the added effects of explicit training on improvement rates. Specifically, with respect to effects on immediate learning (where *n*s started to reach a size that might allow for meaningful interpretation), the effect of the added assignment was significant in improving the *inverse probability* and *replication* fallacies – fallacies that were found to be on average consistently trickier at baseline, across both studies.

Taken together, the two studies in **Project 1** emphasize the instructional merits and strengths of explicit clarification in improving the process of drawing inferences from statistical indices, at least insofar as it concerns the interpretations of conceptual statements, in the context of a MOOC. Beyond explicit clarification, additional insights on how to optimize the instruction of statistical concepts can be gained: For example, imparting a more thorough and nuanced understanding of statistical concepts might be achieved by bringing together (where applicable) certain concepts that share commonalities or differences, and specifically teasing apart how and why they are similar or distinct. In the case of NHST, this can involve introducing the idea of Bayesian inference, or equivalence testing, when explaining the idea behind significance testing (see also **section 5.1.1** below). Lastly, whether differences in accuracy when interpreting significant versus non-significant *p*-value statements stem from i. cognitive biases resulting from top-down pressures, ii. from the fact that the published literature is saturated with statements about significant outcomes, iii. from the possibility that simply some ideas (e.g., why the *clinical or practical significance* fallacy is false) are more intuitive in one context over the other, or iv. from a combination of all three, warrants further investigation. This would be especially relevant for future research that seeks to more practically assess how researchers in their daily careers draw inferences when conducting, evaluating, and reporting theirs and others' real data.

5.1.1. Improving NHST: A practical corollary

In the beginning of the dissertation, the following question was raised: *If p*-value fallacies (or other statistical misconceptions) can be improved, is it possible to determining whether these improvements are a product of overcoming naïve conceptions versus just providing the correct tools to infer the correct meaning – or some combination of both? As demonstrated in **Project 1**, we know that *p*-value fallacies can in fact be improved through the use of explicit clarification and training, so seemingly a lack of information did play a role in why these baseline misconceptions were observed. While the project cannot ultimately speak to whether individuals initially held some deep-seated misconception, or merely lacked the right problem strategy, before instruction successfully improved their understanding, we can nonetheless offer a more practical take-away on how to foster this improvement. The example of improving the *inverse probability* fallacy (see **Box 1**, p. 23) is discussed below.

It is worth first digressing to an explanation of the relationship between *p*-values and effects, emphasizing the frequentist perspective (for further details, see Colling & Szucs, 2018 for recent summary of frequency interpretation of *p*-values, and Morey et al., 2016, for detailed frequency interpretation of CIs). In frequentist terms, evidence of the existence of an effect is inferred as a function of the long run behaviour of a statistical test (*p*-value), across a set of repeated experiments (Neyman & Pearson, 1933). Specifically, the likelihood of observing a small *p*-value is consistent with the existence of a true effect, in the sense that if the null model is true (H_0 = T), statistical tests should only very rarely produce small (significant) *p*-values (i.e. only 5% of the time), whereas if the alternative model is true (H_1 = T), the likelihood over the long run that the test should produce small *p*-values ($p < .05$) is far greater.

Therefore, in some sense, it is easy to see why individuals may fall prey to the *inverse probability* fallacy, i.e. the false belief that a significant *p*-value means that the alternative hypothesis is true, if on some level they are aware of this relationship between a true effect and the greater proportion of significant *p*-values that should be observed over the long run. In other words, researchers may have some intuitions that correctly align with the logic of significance testing, but are misguidedly applying this logic at the level of single *p*-values, rather than across a set of repeated experiments. It follows that in terms of instruction, highlighting this frequentist interpretation may prove most fruitful to eliciting improvements.

Such an approach was in fact taken in week 1 of the MOOC when introducing *p*-values, directly relating the concepts of true versus null effects, statistical power, and expected *p*-value distributions, in order to actively demonstrate this frequentist concept. As exemplified by the plots below (see **Figure 2**), a true effect would be visualized as a more heavily right-skewed distribution of *p*-values (assuming acceptable power, and absence of *p*-hacking or publication bias; see **Figures 2A & 2C**). By contrast, a null effect would be visualized as a flat uniform distribution: Across 100 repeated experiments, each *p*-value should theoretically be observed 1% of the time (see **Figures 2B & 2D**).

What these plots should readily emphasize is that, inherent to the frequentist notion, a *p*-value only becomes meaningful as evidence for an effect when taken in the context of many observations over the long run. Otherwise, a single *p*-value can only provide a statement about the probability of the data, not the probability of a theory or hypothesis. Taken in isolation, a *p*-value affords only partial information about the hypothesis being tested (Ioannidis, 2005; Nuzzo, 2014). Pairing this more dynamic approach of instruction (e.g., use of simulations and/or visualizations) with straight-forward yet theoretically accurate statements, such as *"P*-values are

statements about the probability of the data, not the probability of a theory of hypothesis" (item PV9 of **Study 2**) may also help scaffold and maintain learning.

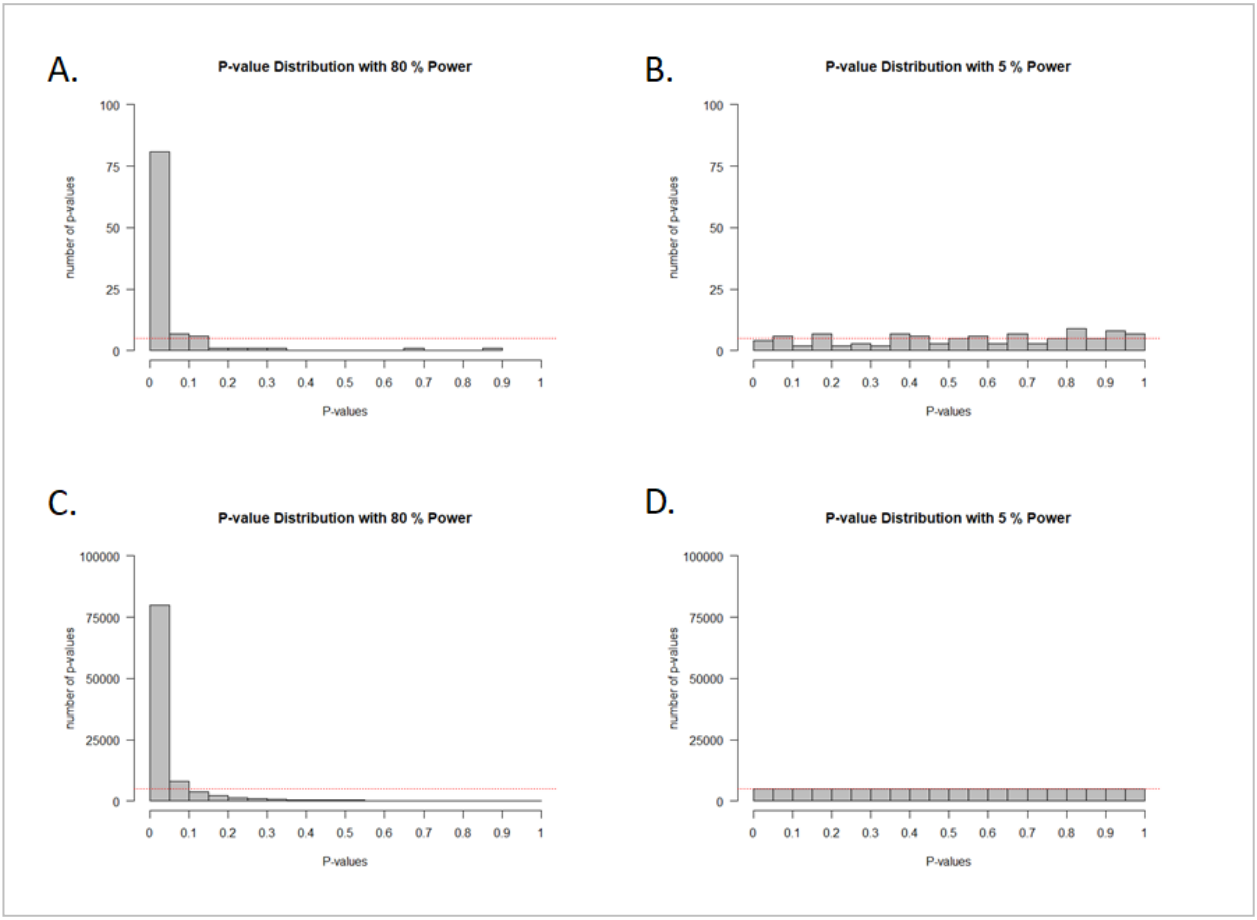

Figure 2. Simulated *p*-value distributions: Theoretical distribution of *p*-values for a one-sample t-test for true vs. null effects; **x-axis:** *p*-values (.0 to 1.0); **y-axis:** raw number of *p*-values in relation to total simulated; red dotted line denotes alpha of 5%. **A. *P*-value distribution with 80% power:** True effect, right-skewed (nSims = 100). **B. *P*-value distribution with 5% power:** Null effect, uniform (nSims = 100). **C. *P*-value distribution with 80% power:** True effect, right-skewed (nSims = 100,000). **D. *P*-value distributions with 5% power:** Null effect, uniform (nSims = 100,000). nSims = number of simulated experiments. R Syntax borrowed from MOOC Assignment 1.1 (Daniël Lakens, *"Improving your statistical inferences"*).

5.2. Project 2: Discussion

Project 2 of the dissertation, **"Exploring indices of repeated k-fold cross-validation as predictors of study replicability"** (Herrera-Bennett, Ong, & Heene), attempted to gain a deeper understanding of the concept and nature of replicability, namely as it concerns how features of individual studies might be used to predict replication success. Specifically, it involved a large-scale re-analysis of the Social Sciences Replication Project (SSRP; Camerer et al., 2018), applying a five-repeated 10-fold cross-validation (CV) technique across 12 of the 21 high-powered replication data sets, in order to compute indices of model generalizability (i.e. average *R*-squared, RMSE, and MAE). These indices were then correlated with the studies' replication success, as indexed by the 'statistical significance criterion', a binary index that categorizes a replication as successful if it is both significant ($p < .05$) and in the same direction as the original effect. The rationale for the study aimed in part to test whether cross-validation, as an index of how a statistical model will generalize to and perform within an independent data set, might theoretically hold in practice when applied to the concept of replicability. Specifically, we presumed that the stronger the model prediction accuracy of the pre-registered replication data sets (i.e. average *R*-squared), the smaller the generalization (or out-of-sample) error (i.e. RMSE and MAE), and in turn the higher the likelihood of replication.

Preliminary results were in line with our predictions, demonstrating a high positive correlation ($r > .80$) between variance explained (average *R*-squared) and replication success, and moderate negative correlations between prediction error indices (RMSE and MAE) and a study's tendency to replicate. Unsurprisingly, the two prediction error indices, though not computationally equivalent, ended up being redundant as predictors, yielding identical Spearman coefficients when correlated with the binary replication indicator. While these results should

necessarily be reassessed after completing the full set of (N = 21) re-analyses, a more comprehensive understanding should also entail investigating the relationship between these CV indices and replicability, by appealing to more than one criterion to judge replication success (i.e. use of a continuous indicator to indicate *degree* of replication). Redundancy of the prediction error indices could also have been affected by having used the binary indicator when running Spearman correlations. Despite these current limitations, one promising outcome of this project's work is, nevertheless, the potential contribution of at least two additional and complimentary statistical indices (i.e. average *R*-squared and RMSE and/or MAE) as predictors of replication, alongside those which have already been pinpointed in previous work. This refers, most prominently, to the observation that *p*-values negatively correlated, and effect sizes positively correlated, with replication success, when evaluated independently in each of the three large-scale replication projects (OSC, 2015; Camerer et al., 2016, 2018). Moreover, when a Random Forest model was trained across a number of the replication projects, including some of the Many Labs projects (see Altjmedt et al., 2019), using over a dozen study features as predictors, model prediction accuracy reached approximately 70%, with the top predictors being the significance level of the *p*-value and the size of the effect. Taken alongside our preliminary findings, this work further corroborates the idea that strength of initial evidence – whether expressed in terms of statistical significance, effect size, or model strength – is an important determinant of study replicability.

What is interesting about the statement above is that there seems to be an intuitive connection between statistical concepts including *p*-value significance and replicability. In fact, as will be discussed in the section below, when researchers were provided the details of the original replication project studies (e.g., OSC, 2015; SSRP, 2018) *before* studies were replicated,

they were fairly accurate in predicting those that would and would not replicate. This is a point worth discussing, especially because from **Project 1**, we know that individuals find the concept of replication (at least as it pertains to *p*-values) fairly difficult (i.e. high baseline *replication* fallacy misconception rates), thus uncovering why replication is more or less intuitive across different contexts would be potentially informative to understand and improve how researchers tend to understand these statistical concepts. The following section addresses this point.

5.2.1. Reconciling replication intuitions: Avenues for future research

Following from the idea above, there seems to exist two seemingly conflicting intuitions about the concept of replication, across researchers. Specifically, it was observed that on the one hand, researchers are apparently intuitive about the replicability of real studies, as evidenced by their fairly accurate ability to predict the future replication success of studies, when provided with information about the original study and replication designs (i.e. prediction markets; Dreber et al., 2015). Yet on the other hand, individuals commonly fall prey to the *replication* fallacy, which is the false belief that a significant *p*-value (e.g., $p = .01$) will directly entail a 99% probability of replicating. The reason why these two statistical intuitions seem particularly conflicting is because in terms of individual study features, *p*-values have systematically been found to be the best predictor of replication success (OSC, 2015; Camerer et al., 2016, 2018; Amrhein et al., 2019).

Therefore, it begs the question as to which single or set of characteristics of a study individuals are appealing to, when making these prediction market judgments, given the overwhelming emphasis and over-reliance there is on *p*-values in the research community. While only speculation, *if* – for argument's sake – individuals are appealing to the size of the *p*-value

when predicting replication, then this would align with their tendency to also fall prey to the *replication* fallacy. Why then is the use of the information accurate (so to speak) in one context, but not in the other? Alternatively, if individuals are *not* singularly using *p*-values as the basis for their predictions, but a combination of information, then why should the relevance of weighting multiple pieces of evidence be seemingly lost on individuals when faced with *replication* fallacy statements? These and other questions warrant further research into how researchers, both novices and experts, reason about replication.

One core distinction which is hard to ignore when assessing both these intuitions is that measures of the *replication* fallacy (e.g., **Project 1**) are delivered in isolation and thus constitute a more theoretical conceptualization about the relationship between *p*-values and the likelihood of an effect to replicate. In contrast, prediction market research assesses researchers' ability to judge the likelihood of a real study replicating, presented not only within the context of the original observed results, but also a study's theoretical rationale and design. In this sense, researchers are technically able to appeal to a host of information, beyond statistical indices, as a basis upon which to make their predictions (i.e. more closely related to **Project 2**). Consequently, where a better grasp of the technical statistical definition of a *p*-value may improve one's performance on *replication* fallacy items, components of effective statistical thinking are likely to play an important role in prediction market accuracy. That is, if we assume that prediction accuracy functions in part by being able to take a more global approach to evaluating evidence, by assessing and optimally weighting multiple pieces of information, then this may help reconcile in part why the shared concept of replication can be misunderstood when measured across different contexts or, put differently, when measured across two different levels of analysis.

Given these two findings, it would be reasonable to speculate that when presented with limited information from which to infer a judgment (e.g., only the *p*-value), some researchers may take for granted properties of other statistical concepts (e.g., statistical power, type-1 error rate) relevant to making a well-founded inference. By contrast, it is likely that others might be better skilled at knowing when to ask these key questions in order to ascertain or realize the extent to which information is unknown or uncertain. In situations where individuals are given more information to work with (e.g., full study details), this could serve to partially close the gap between these two types of thinkers, assuming that the additional pieces of information serves to help scaffold those who would have otherwise neglected to consider them as a relevant to the problem space.

Resolving these types of questions is an active goal among meta-researchers today. Specifically, current collaborations such as the SCORE program (Systematizing Confidence in Open Research and Evidence) and the repliCATS projects, both DARPA-funded (Defense Advanced Research Projects Agency) research projects, specifically bring together interdisciplinary researchers in order to investigate questions related those outlined above. In brief, these collaborations aim to make use of a combination of machine-learning models in conjunction with direct input from a group of experts in order to gain a more comprehensive understanding about which elements of evidence evaluation and reasoning may be playing a role in accurate prediction accuracy. This necessarily goes hand in hand with continuing to identify areas where researchers (experts included) fall prey to misunderstandings, essentially taking the line of research one step further, by investigating how individuals succeed or fail at evaluating statistical concepts when assessed in combination. One fruitful outcome currently underway from **Project 1**, is the contribution of the misconception scales developed and refined in **Studies**

1 and **2** to the repliCATS project, to serve as a subset of baseline controls and discriminators when assessing the reasoning among the group of experts. These and other types of collaborative projects that seek to investigate common questions from distinct perspectives may help pinpoint additional intuitions about statistical concepts among researchers which – correct or misguided – may provide additional information on how to improve instruction.

Taken together, both **Project 1** and **Project 2** provide fairly disparate vantage points from which to consider the questions of *how* researchers evaluate evidence when drawing statistical inferences, and how this process might be better understood and *improved*. That said, when it comes to investigating questions of this complex nature, I would argue that there are merits in gaining a more multi-faceted outlook on such a topic of investigation, if only to contribute to a more cumulative body of research from which to tease out more concrete implications. In fact, upon closer look, not only are there theoretical bridges that can be made between the two works, but such an interdisciplinary approach (e.g., meta-research) may inspire veritable avenues of research to build toward a more comprehensive understanding of the dynamic processes of effective statistical thinking and inference-making. Some general insights and conclusions as an outcome of these two dissertation projects are summarized below.

6. Conclusion

One common theme that has been raised in the current dissertation, and carried throughout the two projects, is the notion of effective statistical thinking and inference-making as a process that involves many moving parts. Therefore, equipping researchers with alternative methods and strategies for drawing inferences has been argued as a goal worth pursuing. With this goal, however, follows the continued need to identify and improve misconceptions or misuses of any

new statistical tool introduced – simply replacing one method with another would be too simplistic an approach. Accordingly, beyond assessing one's understanding of statistical concepts separately, investigating how *multiple* pieces of information are considered collectively and respectively weighted, as well as considering the broader context within which researchers routinely draw inferences (e.g., culture of research), should provide deeper insight into how researchers navigate their problem space. Furthermore, figuring out how this dynamic appraisal of information is more successful among some over others may help uncover how researchers determine which pieces of evidence are most relevant to the specific situation at hand, and may help identify ways to improve this type of flexible thinking.

It follows that if we accept the premise that multiple sources of influence can drive a researcher's process of statistical inference-making, then it would be plausible to assert that there are also multiple means through which to trigger a shift toward an improved approach to drawing statistical inferences. In the same way that it took a crisis at the level of the research community to really start tackling what (NHST) critics have been belaboring for decades, it could be argued that to truly prompt not just immediate but maintained improvements in inference-making among researchers, it demands more than just pointing out misconceptions, but provoking some kind of deeper learning, and/or motivation for change. This might be elicited at the level of conceptual learning (e.g., explicit clarification, insight by comparison), or in terms of a shift in problem strategy (e.g, relaxing constraints on goal-state, or shift toward goal-free problem-solving, like implementation of Registered Reports), or potentially via fundamental changes at the level of the scientific community (e.g., shift in incentive structures).

Finally, another common idea that evolved out of both projects was the idea of statistical intuitions. When it comes to statistics, some concepts will certainly be on average more intuitive

to understand than others which, in terms of implications for instruction, may simply boil down to devoting more time and effort in clarifying them. On the other hand, a more daunting instructional hurdle will be met when statistical intuitions are inherently wrong or misguided, and/or conflict with others which are intrinsically correct. Thus ideally, when teasing apart the complexities of an overarching statistical concept – e.g., confronting the reasons for why one intuition over another may be right versus wrong – it would be arguably advantageous to capitalize upon those common sense notions among novices or experts that intuitively *feel* (and are) correct in order to override, and/or ultimately undo, those that are incorrect. How to achieve this successfully, however, and in practical terms, may be easier said than done. With that said, advances in meta-research and interdisciplinary collaborations seem to be a promising avenue to tackle these sorts of questions.

REFERENCES

(Introduction & General Discussion)

Altmejd, A., Almenberg, A. D., Forsell, E., Ho, T. H., Huber, J., Imai, T., ... & Camerer, C. (2019). Predicting the Replicability of Social Science Lab Experiments.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance.

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*(3), 290-295.

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, *66*(6), 423.

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-16). Dordrecht, The Netherlands: Kluwer academic publishers.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions?. *Journal of Educational Statistics*, *10*(3), 252-268.

Butler, J. (1736). The analogy of religion.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.

Chiesi, F., & Primi, C. (2010). COGNITIVE AND NON-COGNITIVE FACTORS RELATED TO STUDENTS'STATISTICS ACHIEVEMENT. *Statistics Education Research Journal*, *9*(1).

Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for research in mathematics education*, 16-30.

Clement, J. (1987). The Use of Analogies and Anchoring Intuitions to Remediate Misconceptions in Mechanics.

Cohen, J. (1990). Things I have learned (so far). *American psychologist*, *45*(12), 1304.

Cohen, S., Smith, G., Chechile, R. A., Burns, G., & Tsai, F. (1996). Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics*, *21*(1), 35-54.

Colling, L. J., & Szucs, D. (2018, November 2018). Statistical reform and the replication crisis. arXiv:1811.01821

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... & Wilson, S. (2007). Statistical reform in psychology: Is anything changing?. *Psychological science*, *18*(3), 230-232.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall?. *Psychological methods*, *11*(3), 217.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding statistics*, *3*(4), 299-311.

De Vries, R., Anderson, M. S., & Martinson, B. C. (2006). Normal misbehavior: Scientists talk about the ethics of research. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 43-50.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? Perspectives on Psychological Science, 6(3):274–290.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343-15347.

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological monographs*, *58*(5), i.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

Einstein, A., & Infeld, L. (1961). *The Evolution of Physics, Etc.[Edited by Leopold Infeld.]*. Simon & Schuster.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational leadership*, *43*(2), 44-48.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75-98.

Fanelli, D. (2010). "Positive" results increase down the Hierarchy of the Sciences. *PLoS ONE*, *5*(3), 1-10, e10068. doi:10.1371/journal.pone.0010068

Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*(37), 15031-15036. doi:10.1073/pnas.1302997110

Finch, S., & Cumming, G. (1998). Assessing conceptual change in learning statistics. In *In*.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the royal statistical society*, *98*(1), 39-82.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, *1*, 2.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. *The Sage handbook of quantitative methodology for the social sciences*, 391-408.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037-1037.

Gliner, J. A., Vaske, J. J., & Morgan, G. A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife*, *6*(4), 291-301.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual review of psychology*, *40*(1), 631-666.

Greenland, S. (1998). Induction versus Popper: substance versus semantics. *International Journal of Epidemiology*, *27*(4), 543-548.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, 117-159.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1-20.

Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse ofp values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037.

Innabi, H., & Jordan, N. C. H. R. D. (1999). Students' judgment of the validity of societal statistical generalization. In *Proceedings of the international conference on mathematics education into the 21st Century: Societal challenges, issues and approaches*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of evaluation in clinical practice*, *14*(5), 951-957.

Ioannidis, J. P. (2018). Meta-research: Why research on research matters. *PLoS biology*, *16*(3), e2005468.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical trials*, *4*(3), 245-253.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Methodology*, *4*(4), 152-158.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review, 2(3), 196-217. doi:10.1207/s15327957pspr0203_4

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759.

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572-584.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology*.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition*, *25*(6), 1534.

Köhler, W. (1970). *Gestalt psychology: An introduction to new concepts in modern psychology*. WW Norton & Company.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?*(pp. 205-259). Springer, Dordrecht.

Lakens, D. (2019, April 9). The practical alternative to the p-value is the correctly used p-value. https://doi.org/10.31234/osf.io/shm8v

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?. *Journal of the American statistical Association*, *88*(424), 1242-1249.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 293-337.

Lyu, Z., Peng, K., & Hu, C. P. (2018). P-value, Confidence Intervals and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*, 868.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, *1*(2), 161-175.

Maier, N. R. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, *12*(2), 181.

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature*, *435*(7043), 737-738. doi:10.1038/435737a

Maslow, A. H., & Wirth, A. G. (1966). The psychology of science: A reconnaissance.

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In *Optimality* (pp. 77-97). Institute of Mathematical Statistics.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Morey, R. D., Romeijn, J.-W., and Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. Journal of Mathematical Psychology, 72:6–18.

Newell, A., & Simon, H. A. (1975). Computer science as empirical inquiry: Symbols and search. *PHILOSOPHY OF PSYCHOLOGY*, 407.

Neyman, J. (1955). *The problem of inductive inference*. Interscience.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289-337.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, *506*(7487), 150.

Oakes, M. W. (1986). *Statistical inference*. Epidemiology Resources.

Ohlsson, S. (2012). The problems with problem solving: Reflections on the rise, current status, and possible future of a cognitive research paradigm. *The Journal of Problem Solving*, *5*(1), 7.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530.

Piaget, J. (1975). *L'équilibration des structures cognitives: problème central du développement* (Vol. 33). Presses universitaires de France.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education*, *66*(2), 211-227.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. *What if there were no significance tests*, 335-391.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, *61*(4), 293-316.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Smith III, J. P., Disessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The journal of the learning sciences*, *3*(2), 115-163.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98-113.

Spence, J. R., & Stanley, D. J. (2018). Concise, Simple, and Not Wrong: In Search of a Short-Hand Interpretation of Statistical Significance. *Frontiers in Psychology*, *9*, 2185.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, *55*(1), 84-94.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.

Vacha-Haase, T., & Ness, C. M. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice*, *30*(1), 104.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, *14*(5), 779-804.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Declaration for Authorship
+
Academic Integrity Statement

I, **Arianne Herrera-Bennett**, declare that this dissertation *"How do researchers evaluate statistical evidence when drawing inferences from data?"* is my original academic work, and that it was not presented, and will not be presented to any other university than LMU for other or a similar doctoral degree award.

I declare that this is my original academic work and has not been copied from other works, and that other authors' ideas have been cited properly as for any scholarly work.

Arianne Herrera-Bennett

Munich. 03. 06. 2019

[end = TRUE]

;)