The statistical crisis in science: How is it relevant to clinical neuropsychology?∗

Andrew Gelman† and Hilde M. Geurts‡

7 June 2016

Abstract

There is currently a statistical (and replication) crisis in science. Social psychology has been at the heart of this crisis, but the lessons learned are relevant for other fields. We discuss three examples of replication challenges and some proposed solutions, and then consider the applicability of these ideas to clinical neuropsychology. In addition to procedural developments such as preregistration and open data and criticism, we recommend that data be collected and analyzed with more recognition that each new study is a part of a learning process. The goal of improving neuropsychological assessment, care, and cure is too important to not take good scientific practice seriously.

1. Introduction

In the last few years psychology has been rocked with what has been called a replication crisis. Recent theoretical and experimental results have cast doubt on what previously had been believed to be solid findings in social psychology.

From the theoretical direction a close examination of the properties of hypothesis testing has made it clear that conventional “statistical significance” does not necessarily provide a strong signal in favor of scientific claims (Ioannidis, 2005, Simmons, Nelson, and Simonsohn, 2011, Francis, 2013, Button et al., 2013, Gelman and Carlin, 2014). From the empirical direction, several high-profile studies have been subject to preregistered replications that have failed (see, for example, Doyen et al., 2012, and Open Science Collaboration, 2015). This all creates challenges for drawing conclusions regarding behavioral and cognitive theories, and in mapping research findings to clinical practice. In the present paper we shall first review what is known as the replication crisis within social psychology and then pose the question of its relevance to clinical neuropsychology and clinical practice, arguing that from this crisis lessons can be learned that are of importance to a wide range of research fields, including clinical neuropsychology.

2. The replication crisis in social psychology

Lately there have been various discussions regarding a wide range of psychological studies, but we will use three different examples to illustrate different aspects of the replication crisis.

The first example is that specific research groups have been promoting the idea that holding an open posture—the “power pose”—can “cause neuroendocrine and behavioral changes” so that you “embody power and instantly become more powerful.” This claim has received lots of publicity, but its specifics have been questioned both by statistical arguments detailing how the published

∗For The Clinical Neuropsychologist. We thank the U.S. National Science Foundation, Institute of Education Sciences, Office of Naval Research, and the Netherlands Society for Scientific Research for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡University of Amsterdam, Department of Psychology, Brain and Cognition Section, Dutch ADHD and Autism Research Center, Amsterdam, The Netherlands, and Dr. Leo Kannerhuis, Department of Research, Development, and Innovation, Doorwerth, The Netherlands.
findings could have occurred by chance alone (data exclusion, interactions, choices in regressions, miscalculation of p-values) and by a failed replication that was larger in size than the original study (Ranehill et al., 2015, Simmons and Simonsohn, 2015). From a science-communication perspective, a major problem is that the researchers on the original study (Carney, Cuddy, and Yap, 2010) seem to refuse to accept that their published findings may be spurious, despite the failed replications, the statistical theory that explains how they could have obtained low p-values by chance alone (Gelman and Loken, 2014), and the revelations of errors in their calculations (Gelman, 2016a). Of course in science mistakes can be made, but the crucial aspect of science is that we recognize these mistakes and learn from them so we can increase the body of knowledge regarding a specific topic and know which findings are actually false. Replication—and learning from replications both successful and failed—is a crucial aspect of research.

For our second example, a clear demonstration of the value of replication was discussed by Wagenmakers et al. (2015): “the phenomenon of social priming, where a subtle cognitive or emotional manipulation influences overt behavior. The prototypical example is the elderly walking study from Bargh, Chen, and Burrows (1996); in the priming phase of this study, students were either confronted with neutral words or with words that are related to the concept of the elderly (e.g., ‘Florida’, ‘bingo’). The results showed that the students’ walking speed was slower after having been primed with the elderly-related words.” Noted cognitive psychologist Daniel Kahneman wrote of such studies: “the reaction is often disbelief ... The idea you should focus on, however, is that disbelief is not an option. The results are not made up, nor are they statistical flukes. You have no choice but to accept that the major conclusions of these studies are true.”

Maybe not. Wagenmakers et al. continue: “At the 2014 APS annual meeting in San Francisco, however, Hal Pashler presented a long series of failed replications of social priming studies, conducted together with Christine Harris, the upshot of which was that disbelief does in fact remain an option.” The quote from Kahneman is, in retrospect, ironic given his celebrated work with Tversky from the 1970s, finding that research psychologists were consistently overstating the statistical evidence from small samples. In the example of social priming, the unsuccessful replications performed by outside research groups were important, even (or especially) for an apparent finding that had already found its way into psychology textbooks. Being convinced that something is a genuine effect is not sufficient.

There has also been pushback at revelations of statistical problems in published work. For example, in his response to the failed replications of his classic study, Bargh (2012) wrote:

There are already at least two successful replications of that particular study by other, independent labs, published in a mainstream social psychology journal... Both appeared in the Journal of Personality and Social Psychology, the top and most rigorously reviewed journal in the field. Both articles found the effect but with moderation by a second factor: Hull et al. 2002 showed the effect mainly for individuals high in self consciousness, and Cesario et al. 2006 showed the effect mainly for individuals who like (versus dislike) the elderly... Moreover, at least two television science programs have successfully replicated the elderly-walking-slow effect as well, (South) Korean national television, and Great Britain’s BBC1. The BBC field study is available on YouTube.

This response inadvertently demonstrates a key part of the replication crisis in science: the researcher degrees of freedom that allow Bargh to declare victory in so many ways. If the papers by Hull and Cesario had found main effects, Bargh could have considered them to be successful replications of his earlier work. Instead he considered the interactions to represent success—but there are any number of possible interactions or subsets where an effect could have been found.
Realistically there was no way for Bargh to lose: this shows how, once an idea lodges in the literature, its proponents can keep it alive forever (and can promote it on television, presenting speculation as if it were settled work). This works against the ideally self-correcting nature of science.

Our third example comes from Nosek, Spies, and Motyl (2012) who performed an experiment in which,

Participants from the political left, right and center ($N = 1,979$) completed a perceptual judgment task in which words were presented in different shades of gray. Participants had to click along a gradient representing grays from near black to near white to select a shade that matched the shade of the word. We calculated accuracy: How close to the actual shade did participants get? The results were stunning. Moderates perceived the shades of gray more accurately than extremists on the left and right ($p = .01$). Our conclusion: political extremists perceive the world in black-and-white, figuratively and literally. Our design and follow-up analyses ruled out obvious alternative explanations such as time spent on task and a tendency to select extreme responses.

This might seem a legitimate observation and could have been a finding when published as such would gained a lot of media attention. Luckily before submitting to a journal, the researchers themselves performed a replication, reporting that they “ran 1,300 participants, giving us .995 power to detect an effect of the original effect size at $\alpha = .05$.” The result:

The effect vanished ($p = .59$).

This is an inspiring example of good science, where researchers applied a stringent critique to their own work. It is good practice that other research labs try to replicate published work, but trying to replicate your own findings first is an important first step.

These and other examples have made it clear to many observers, inside and outside the field of psychology, that top journals repeatedly publish and promote claims based on the flimsiest of evidence. Similar problems have been noted in neuroimaging (Vul et al., 2009) when the data analysis is selected after the data have been collected, $p$-values cannot be taken at face value.

The above examples demonstrate both the problems that can arise with such $p$-values (namely, publication of results that ultimately do not replicate) and the different ways that researchers can address this concern, ranging from denial (as with the power-pose researchers) to self-examination (as in the “50 shades of gray” study). The reaction to the elderly walking study demonstrates the role of defaults in the reception of published work. Kahneman showed the attitude of deferring to the literature, while Wagenmakers et al. demonstrate the virtue of skepticism.

Assessing the strength of evidence is difficult, and in the context of interpreting research there can be a tendency to count the same evidence multiple times. Suppose a research finding is statistically significant and the estimated effect size is large and it is backed up by theory and it is published in a major journal and there is follow-up literature of successful conceptual replications. This might seem like several pieces of evidence confirming the truth of the theory, but in a sense this could really be just one piece of information appearing repeatedly: Only findings that are statistically significant will be published; in an experiment with small sample size and high variability, any statistically significant pattern will necessarily be large; when theories are vague enough that they can explain any observed pattern, positive or negative; it is no surprise that a noteworthy finding will be followed up; and given the flexibility of theory and researcher degrees of freedom in data processing, analysis, and presentation, future researchers should have no difficulty obtaining statistically significant results that can be construed as being part of the general constellation of findings that are consistent with the theory.
This “sociology” of the process of research and publication completes the picture: it explains how well-meaning researchers can perpetuate a subfield for decades, even in the absence of any consistent underlying effects; see Smaldino and McElreath (2016) for a simple agent-based model. Which brings us to our central question here: To what extent can this “false-positive psychology” (in the words of Simmons, Nelson, and Simonsohn, 2011) also affect clinical neuropsychology, and how can we expect these fields to change now, given our new found awareness of these problems?

3. **General solutions in response to the replication crisis**

   The lessons learned from the crisis in social psychology should be relevant for many other research fields. Many of the most publicized studies that have failed to replicate already seem a bit silly to many people and may seem to have little relevance to clinical research or practice. We do believe, however, that the recent wave of revelations regarding junk psychological science has effects on how to think about research in clinical neuropsychology. As long as effects are variable and measurements are noisy, inferences based on p-values can mislead.

   Statistical significance is a lot less meaningful than traditionally assumed, for two reasons. First, abundant researcher degrees of freedom (Simmons, Nelson, and Simonsohn, 2011) and forking paths (Gelman and Loken, 2014) assure researchers a high probability of finding impressive p-values, even if all effects were zero and data were pure noise. Second, as discussed by Gelman and Carlin (2014), statistically significant comparisons systematically overestimate effect sizes (type M errors) and can have the wrong sign (type S errors).

   The various ideas that have been proposed for resolving the replication crisis can be characterized as follows:

   - **Science communication:** Not restricting publication to “statistically significant” results; outlets for publishing replication attempts; collaborations between disagreeing researchers.

   - **Design and data collection:** Preregistration; design analysis using prior estimates of effect size; more attention to accurate measurement; replication plans baked into the original design.

   - **Data analysis:** Bayesian inference; hierarchical modeling of multiple outcomes and potential explanatory factors; meta-analysis; control of error rates.

   These ideas interact in different ways. For example, consider the idea of paying more attention to accurate measurement, which takes on new value in the context of abandoning statistical significance. Under the current system, researchers may have little motivation to improve their measurements: if they have achieved \( p < .05 \), then they can feel that their measurements have already succeeded, so why worry? But when researchers are aware of the problems of forking paths, and magnitude and sign errors, the benefits to more accurate measurements are clearer.

4. **Relevance to clinical neuropsychology and clinical practice**

   We can envision various ways in which the statistical crisis in science can be affecting the understanding and practice of clinical neuropsychology. While many clinicians were probably already well aware of difficulties in translating research findings to clinical practice, the recent debates regarding the replication problems might have strengthened them in the idea that much scientific research is not directly relevant for their daily work. This might be disadvantageous and even dangerous, in that inconvenient findings that might be important to clinical practice could be disregarded as “just
another study.” Hence, it is crucial that those who give guidance to clinicians know how to evaluate different sorts of statistical evidence.

For example, there could be studies recommending particular therapies, studies published in top journals with statistically significant results, but which for the reasons discussed above should not be trusted. Are there such studies? We could also flip the question around: instead of asking which particular findings or areas of research are suspect, we could ask whether any statistically-based studies can be trusted.

We have not systematically examined the field of clinical neuropsychological research; rather, we give a series of examples to illustrate the (related) challenges in translating research to clinical practice and in dealing with the aforementioned statistical issues. We try to illustrate how clinicians can distinguish the good from the bad and what researchers did or should have done to ensure that the obtained findings actually make sense.

Within clinical neuropsychology it is standard practice to use a series of neuropsychological tests for two purposes: normative assessment of a patient’s specific strengths and weaknesses across a wide range of cognitive domains, and serial assessment of changes in cognition. In combination with findings from adjacent disciplines, the individual cognitive profile is then often used as part of determining a plausible diagnosis, to select a suitable intervention, and to evaluate an intervention. In scientific clinical neuropsychological research, case studies as well as group studies are common. Difficulties arise when one wants to use findings from either sort of research study and translate to clinical practice. With case studies it is hard to determine whether findings can be generalized to other cases as one needs to determine when a case is similar enough to the cases from the study. With group studies it is hard to determine whether the observed effect is sufficiently large to have clinical implications for each individual of a specific group.

Five major challenges arise:

The first challenge is that clinical neuropsychological tests can include a wide range of outcomes. For example, the Wisconsin Card Sorting Test (WCST) is an executive function task with fourteen different dependent measures. This in itself is not a problem for research, but it will be a problem when one wants to test the idea that a specific patient group performs worse than other groups on this task when the researcher did not determine ahead of time which dependent measure is the crucial measure to test this idea. One can observe no differences between the groups of interest on several of the measures (for example, total number of cards used, failure to maintain set, and perseverative errors) but also, just by chance, a difference on one specific measure (for example unique errors). If a researcher already has strong beliefs about what to expect or if journals are predisposed to publish positive findings, there is the risk that in the paper only the dependent measure with the success will be discussed and the rest will be ignored—and, more generally, the problem of overinterpretation of differences between statistical significance and nonsignificance (Gelman and Stern, 2006). It is in itself not a problem to focus on a subset of outcomes, but this should be motivated by a strong prior theoretical argument. One way to demonstrate a prior theoretical commitment is preregistration, so that it is clear which part of the study is confirmatory and which part is exploratory. More generally, though, we recommend analyzing all outputs using graphical displays to show a grid of estimates without focusing on statistical significance, and using multilevel models to partially pool information (Gelman, Hill, and Yajima, 2012). When performing meta-analysis it is important to use all the information from each study. A meta-analysis performed using only published summaries can be biased given that each study will tend to record the subset of analyses that are statistically significant, thus leading to a systematic overconfidence and overestimation of effect sizes (Button et al., 2013). As guideline for a specific type of observational study, researchers can use STROBE (www.strobe-statement.org); and for meta-analyses and reviews, we refer to PRISMA (http://prisma-statement.org/)
The second challenge is replicating findings in specific and sufficiently large patient groups. Replication is of importance in all fields of science, but is more challenging in the human sciences, where effects vary across people and situations, measurements are noisy and often indirect, and where there can be serious concerns extrapolating from experimental to routine clinical settings. When conducting studies in first year psychology students one can run the same experiment multiple times with independent samples. However, when studying rare disorders this is much harder. Still replication is crucial. Collaborations across labs and countries can aid in this endeavor, but also when setting up a new study one could can combine this with a replication study.

The third challenge is determining if a finding is considered robust and should be implemented in clinical practice. For example, the evidence pyramid for behavioral rehabilitation interventions often starts with the observation that some patients seem to benefit from a specific intervention. Multiple case series studies can provide some first evidence on whether this observation can be replicated in a more structured setup. Not just medical treatment trials can be registered in databases such as www.clinicaltrials.com and other country-specific websites. Here one needs to preregister the study design and the primary and secondary outcome measures of the study. We recommend that both academic researchers and clinical researchers to use these preregistration opportunities. Moreover, we recommend that researchers take notice of existing guidelines for treatment studies (see, for example, http://www.consort-statement.org/). However, this does not resolve the question of when there is sufficient evidence to recommend adopting a specific treatment. Here we think it would make sense to perform a decision analysis, explicitly balancing costs, risks, and benefits. We are well aware that sufficient information may not be at hand to perform such decision analysis, but we feel that such calculations are valuable even if their main role is to reveal what key decision-relevant information remains unavailable.

The fourth challenge goes in the other direction: how can we determine when there is sufficient evidence to disregard earlier conclusions that already are widespread in clinical practice? For example, for decades the dominant model of Alzheimer’s disease was the amyloid cascade hypothesis involving a specific temporal ordering in some biomarkers (Jack et al., 2013). However, findings from a meta-analysis cast doubt on this model (Schmand, Huizenga, and van Gool, 2010). In the six years since its publication, this paper has not seemed to have had a serious impact on the field. How much we can believe a research finding, how much replication are needed to counterbalance methodological criticisms and unsuccessful replications, and when evidence is sufficient and sound enough to transfer knowledge to the clinic?

The fifth challenge is to translate group findings to individual patients. In various journals, researchers are asked to describe the clinical implications of their findings. While it is good to discuss this, there is a risk of overstating what has been learned. Even when there is overwhelming evidence of an average differences, is can be hard to determine what this means for an individual patient. Conversely, how can we determine the extent to which an individual pattern of strengths and weaknesses is in line with research findings that apply to groups of patients? There are large initiatives such as ANDI (http://www.andi.nl/home/) where individual profiles are tested against large combined data sets, in which statistical methods are already implemented so the users can use these statistical techniques without knowing all the details of these techniques. This is an example of how collaboration between different fields in combining existing datasets can lead to a tool that is directly relevant for clinicians and patients.

5. Example: cognitive training for working memory

For an example of the challenges in assessing evidence for clinical practice, consider computerized cognitive training programs for working memory training. The first published findings showed pos-
itive effects for children with ADHD (Klingberg, Forssberg, and Westerberg, 2002). Two questions arise: first, whether working memory capacity can actually be enhanced by cognitive training; second, whether this will have an effect on symptomatology as some argued that a working memory problem could be a primary deficit in ADHD. To address both questions, the authors presented two small studies, one with children with an ADHD diagnosis ($N = 7$ per treatment group; the intervention versus an active control condition) and one in adults ($N = 4$, with no control group). In both studies the authors observed positive treatment effects. After this proof-of-concept study the authors conducted a larger randomized clinical trial with 26 children with ADHD in the intervention compared to an active control and again observed positive effects of the tested treatment (Klingberg et al., 2005). Since then a plethora of studies from the same and other research groups have appeared trying to replicate the findings with an adjusted version of the same training. Moreover, Klingberg set out to conduct research to study the underlying mechanism to understand why the intervention might work. So far so good, but does this imply that there is overwhelming evidence in favor of this intervention? Can you tell people with ADHD that this intervention is useful? Do other similar interventions have similar effects? Does it matter whether someone has an ADHD diagnosis or is it beneficial for all of us? Both positive and negative replications and meta-analyses have been published (for example, Buschkuehl and Jaeggi, 2010, Chacko et al., 2013, Evans et al., 2013, Melby-Lervag and Hulme, 2013, Morisson and Chein, 2011, Rabipour and Raz, 2012, Rapport et al., 2013, Shipstead et al., 2012, and Spencer-Smith and Klingberg, 2015). The evidence in favor and against the claim seems to be weighted differently in these different reviews. It is exactly this what might confuse clinicians (and patients) in deciding how to weight the evidence.

Moreover, this working memory training has been commercialized (CogMed) based on the evidence in favor of effectiveness (however the original researcher is currently not involved in this commercial enterprise), and when clinicians are trained to use this training the critical side is hardly communicated. Recently there was a lot of positive press for CogMed based on a recent meta-analysis by Spencer-Smith and Klingberg with the conclusion was that working memory training had a medium positive effect on attention in daily life and therefore has clear clinical relevance. However, the estimate was much lower when others reanalyzed the same data after fixing coding errors were made and correcting for publication bias (Dovis et al., 2015). What was left was just a small effect which does not warrant the conclusion that there is a clear clinical relevance. Such a correction gains less media coverage and shows why it is important to go back to the literature and remain critical about what you read.

In order to reduce this kind of confusion, different disciplines work together to develop clinical guidelines. Typically different stakeholders are involved who together need to translate findings from both standardized evaluations of literature and clinical practice into pragmatic suggestions for the clinical field. This is not an easy endeavor and still conclusions can be subjected to debate but at least clinicians who read the guidelines would be up to date with the scientific pros and cons of specific assessment measures and interventions. Therefore, we encourage clinicians to check whether there are clinical guidelines or specific Cochrane reviews regarding a specific topic.

6. Discussion

All of us, researchers and laypeople alike, have our own observations and ideas about how communication works, how memory works, how information is perceived, and when we should consider something a disorder or not. This makes the field of psychology sensitive to opinions and selective reading of the literature. In the world of pop psychology, and regretfully also in publications in top journals human behavior can seem deeply irrational as, for example, presidential elections are said to be determined by the outcomes of college football games, women’s votes are apparently
determined by their time of the month, men’s attitudes toward economic redistribution are affected by upper-body strength, exposure to the subliminal image of a smiley face has large effects on attitudes toward immigration, the choice of male or female name for a hurricane affects disaster response, and so on. All these claims were promoted in respected journals or media outlets, and all have been called into question on methodological grounds (see Gelman, 2015). If you were to believe all or even many of these effects, you would have to believe in a capricious world in which tiny, often imperceptible stimuli have huge impacts on our attitudes and behaviors. Conversely, a downgrading of belief in all these effects implies a more rational, predictable world, perhaps more amenable to psychological therapies. So even though it might seem not a serious issue as the aforementioned “findings” might not affect policy makers or change our own behavior, we do need to take this seriously as it reduces the trust in the coherence of science.

Although, awareness of the statistical crisis in psychology research is not new (see, for example, Meehl, 1990, and he had been writing about this for decades by then) it is now in the center of attention. Recent papers by Ioannidis (2005), Simmons, Nelson, and Simonsohn (2011), Open Science Collaboration (2015), and others have raised the awareness that it is time for action. This action should not be restricted to the fields that happened to be in the center of attention but should will translate across disciplines to raise awareness about good research practices.

In recent years we have come to realize that many seemingly solid results in experimental psychology can be explained as nothing but creative interpretation of small-sample variation. There is a trust gap in psychology research and also in other fields such as medicine which feature the combination of small samples, highly variable effects, and an intense pressure to obtain statistically significant results. There have been some high-profile failed replication attempts and a call to replicate studies more generally. Laboratory studies or online surveys are easy enough to replicate. But for clinical research, replication is more challenging for reasons both of cost and ethics. So, yes, we should think of replication as a standard—we should always ask if we think a finding would replicate—but in many cases this standard is theoretical, and we must perform inference about replications using what we call “design analysis” (Gelman and Carlin, 2014) to get some estimate of the extent to which published results can overestimate underlying effect sizes and even go in the wrong direction.

The goal of improving care and cure is too important to wait until all the statistical problems associated with publication bias have been fixed. We need to analyze previous and existing data more comprehensively, perform directly replications of studies of interest, and design new studies with replication in mind, moving away from the idea of statistical significance as a demonstration of effectiveness and instead thinking of each new study as part of a learning process.

References


