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

Gelman, A.; Geurts, H.M.

DOI
10.1080/13854046.2016.1277557
Publication date
2017
Document Version
Final published version
Published in
The Clinical Neuropsychologist

THE CLINICAL NEUROPSYCHOLOGIST, 2017
VOL. 31, NOS. 6–7, 1000–1014
https://doi.org/10.1080/13854046.2016.1277557

Routledge
Taylor & Francis Group

CE The statistical crisis in science: how is it relevant to clinical neuropsychology?

Andrew Gelman[a,b] and Hilde M. Geurts[c,d]

[a]Department of Statistics, Columbia University, New York, NY, USA; [b]Department of Political Science, Columbia University, New York, NY, USA; [c]Dutch ADHD and Autism Research Center, Department of Psychology, Brain and Cognition Section, University of Amsterdam, Amsterdam, The Netherlands; [d]Dr. Leo Kannerhuis, Department of Research, Development, and Innovation, Doorwerth, The Netherlands

ABSTRACT

There is currently increased attention to the statistical (and replication) crisis in science. Biomedicine and social psychology have been at the heart of this crisis, but similar problems are evident in a wide range of fields. We discuss three examples of replication challenges from the field of social psychology 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.

ARTICLE HISTORY
Received 9 June 2016
Accepted 22 December 2016

KEYWORDS
Replication crisis; statistics; sociology of science

1. Introduction

In the last few years psychology and other experimental sciences have 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 various fields (e.g. see Baker, 2016; Begley & Ellis, 2012; Button et al., 2013; Dumas-Mallet, Button, Boraud, Munafo, & Gonon, 2016; Ioannidis & Panagiotou, 2011). Here we focus on psychology, a field in which big steps have been made to try to address the crisis.

Where do we refer to when we speak of the statistical crisis in science? 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 (Button et al., 2013; Cohen, 1994; Francis, 2013; Gelman & Carlin, 2014; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). From the empirical direction, several high-profile studies have been subject to preregistered replications that have failed (see e.g. Doyen, Klein, Pichon, & Cleeremans, 2012; Open Science Collaboration, 2015). This all creates challenges for drawing conclusions regarding psychological theories, and in mapping research findings to clinical practice.

CONTACT Andrew Gelman gelman@stat.columbia.edu

In the present paper we shall first review the crisis within social psychology and then pose the question of its relevance to clinical neuropsychology and clinical practice. We do not attempt to estimate the extent of replication problems within the field of clinical neuropsychology; rather, we argue that from the general crisis in the social and biological sciences, lessons can be learned that are of importance to all research fields that are subject to high levels of variation. Proposed solutions are relevant for clinical neuropsychological research, but for some specific challenges there are not yet any clear remedies. In line with *The Clinical Neuropsychologist*'s record of scrutiny (e.g. Schoenberg, 2014), we hope to keep clinicians and researchers alert when reading papers and designing new studies.

2. The replication crisis in social psychology

We shall use three different examples from social psychology to illustrate different aspects of the replication crisis.

The first example is a claim promoted by certain research groups, and widely promulgated in the news media, that holding an open posture – the 'power pose' – can 'cause neuroendocrine and behavioral changes' so that you 'embody power and instantly become more powerful.' These and related assertions have been questioned both by statistical arguments detailing how the published 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 & Simonsohn, 2015). From a science-communication perspective, a major challenge is that the researchers on the original study (Carney, Cuddy, & 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 & Loken, 2014), and the revelations of errors in their calculations (e.g. 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.

As our second example, we use a clear demonstration of the value of replication focusing on the phenomenon of social priming, where a subtle cognitive or emotional manipulation influences overt behavior. The prototypical example for social priming is the elderly walking study from Bargh, Chen, and Burrows (1996). Students were either confronted with neutral words or with words that are related to the concept of the elderly (e.g. 'bingo'). This is the so-called priming phase which seemed to be effective as students' walking speed was slower after having been primed with the elderly-related words. Wagenmakers, Wetzels, Borsboom, Kievit, and van der Maas (2015) critically discuss these findings and quote cognitive psychologist Kahneman who 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 such as reported in Tversky and Kahneman (1971), finding that research psychologists were consistently overstating the statistical evidence from small samples. Being convinced that something is a genuine effect is not sufficient, and can sometimes even trick thoughtful researchers into believing in a false conclusion. Criticism and replication are essential steps in the scientific feedback loop.

There has, of course also been push back 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 (vs. 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 allows a researcher to declare victory in so many ways.

To stay with this example: 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 many 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. This works against the ideally self-correcting nature of science. We are not saying here that replications should only concern main effects; rather, the purported replications reported by Bargh are problematic in that they looked at different interactions in different experiments justified with *post hoc* explanations.

Our third example comes from Nosek, Spies, and Motyl (2012) who performed an experiment focusing on how political ideology related to performance on a perceptual judgment task. Their initial findings were, to use their own words, 'stunning.' Based on the results the conclusion could have been that political extremists perceive the world in black-and-white, figuratively and literally. In their paper they state: '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. Fortunately, 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, especially in this sort of lab setting where replication requires little effort.

These and other examples have made it clear to many observers, inside and outside the field of psychology, that some journals repeatedly publish and promote claims based on the flimsiest of evidence. Similar problems have been noted in, for example, neuroimaging:

when the data analysis is selected after the data have been collected, *p*-values cannot be taken at face value (Vul, Harris, Winkielman, & Pashler, 2009). The given 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 perceptual judgment study). The reaction to the elderly walking study demonstrates the role of defaults in the reception of published work. One should not defer to the literature, but rather examine published results in the context of their data, methods, and theoretical support.

Even though there are more and more practical guidelines available to help clinicians and researchers to evaluate different types of studies (see e.g. http://www.cebm.net/critical-appraisal for practical guidelines), assessing the strength of evidence remains difficult. In the context of interpreting research there can be a tendency to count the same evidence multiple times. For example, a research finding might be 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. But 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. This raises the question to what extent, even though clinical researchers often are (or should be) trained to be aware of these problems, can this 'false-positive psychology' (in the words of Simmons et al., 2011) still affect clinical neuropsychology? Recent events have revealed some systemic problems in social science research, but this crisis has led to a series of positive developments that we hope will eventually lead to a more solid foundation.

3. General solutions in response to the replication crisis

The lessons learned from the crisis in social psychology are relevant for many other research fields. Many of the studies that gained a lot of media attention and have failed to replicate might already seem a bit silly to many people, but the crisis is by no means unique to social psychology – the same problems have been noted in neuroimaging, medicine, economics, political science, and just about every other field involving statistical analysis on human data – and it would be a mistake to suppose it is not relevant for clinical neuropsychology as well. 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 (see Cohen, 1994; Meehl, 1978), for two reasons. First, abundant researcher 'degrees of freedom' (Simmons et al., 2011) and 'forking paths' (choices in data processing and analysis that are contingent on data; Gelman & 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). Statistically significant estimates must by construction be at least two standard errors from zero, thus a selection effect which biases effect-size estimates upward, often severely, by a factor of two to ten or more. However, a graphical depiction of reliability intervals, for example, might help to interpret reported *p*-values (see e.g. Cumming & Finch, 2005), even if it will not solve the aforementioned problems of the overinterpretation of statistical significance.

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

- *Science communication*: Not restricting publication to 'statistically significant' results; publication of replication attempts, both positive and negative; collaborations between disagreeing researchers; detailed methods sections in papers so that outsiders can actually try to replicate a study.
- *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.

Variants of these ideas have been proposed by many observers, in recognition that non-replication is a crisis of the social process of science, not just a set of individual errors. To put it another way, suppose that, in an instant, we could eliminate the 'questionable research practices' of hiding of null results, unacknowledged multiple paths in data analysis, and *post hoc* hypothesis testing. Even in this case, we would still be left with a seriously flawed paradigm of scientific discovery in which researchers test their ideas with low-power, hard-to-replicate studies and are too often rewarded for spurious, but exciting, results. Conversely, reforms to the social process of science seem unlikely to result in much improvement if not accompanied by a better understanding of statistical evidence. Hence one needs to push to reform scientific measurement, data analysis, communication, and the incentives to scientists, all in concert.

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 (which allow statistically significant *p*-values to be routinely extracted even from pure noise) and of the way in which selection for statistical significance leads to overestimates of effect sizes, the benefits to more accurate measurements become more clear. This is why researchers in a wide range of fields, including clinical neuropsychology, stress the importance of statistics and measurement when designing studies and when interpreting research findings.

4. Relevance to clinical neuropsychology and clinical practice

There is a good foundation of research advice in clinical neuropsychology on which to build. For example, Millis (2003) already spoke of the seven deadly sins of statistical practices we should avoid in our research; also see, for example, Cohen (1994), Bezeau and Graves (2001), Schatz, Jay, McComb, and McLaughlin (2005), and Button et al. (2013). The problems addressed by these scholars are related to what have been discussed above, remain highly relevant, and can indeed be extended. For example, Millis warned of uncorrected multiple comparisons, a problem that is exacerbated by multiple *potential* comparisons (Gelman & Loken, 2014), and his recommendation of power analysis can be amended by warning of the danger of overconfidence when power analysis is performed based on published effect size estimates, which tend to be overestimates. Some potential good news is that in clinical neuropsychology we often deal with larger effect sizes and more careful measurement than in some of the least replicable areas of experimental social psychology. But, even though several of the solutions are by no means new, it seems we all need constant reminders to ensure that we actually use the lessons learned. This is the major goal of the current paper, to connect many different threads of good research practice.

We can envision various ways in which the statistical crisis in science can be affecting the understanding and practice of clinical neuropsychology. The field has its roots in clinical practice which has resulted in a stream of publications focusing on how to conduct scientific studies which can be translated to the daily work of practitioners; see, for example, Hageman and Arrindell (1999), Graham and Tetroe (2007), Strauss, Sherman, and Spreen (2006), Duff (2012), Lezak, Howieson, Bigler, and Tranel (2012), and Shabbir and Sanders (2014). However, even when many clinicians and researchers are already well aware of difficulties in application of research findings to clinical practice, recent debates regarding the replication crisis might have strengthened some clinicians' doubt about the relevance of much scientific research to the real world. 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 earlier in this paper 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. The latter is the angle we take in the current paper. Rather than attempting to systematically examine the field of clinical neuropsychological research, we have collected some examples to illustrate the (related) challenges in translating research to clinical practice and in dealing with the aforementioned statistical issues.

We discuss several pitfalls, solutions, and continuing challenges within the field of clinical neuropsychology, but there are many findings that have replicated. Consider, for example, dissociation between perceiving color vs. movement (Zeki, 1990; Zihl & Heywood, 2016), dichotic listening (Hugdahl, 2011), the typical clock drawings of neglect patients (Chen & Goedert, 2012), the effect of passing time for memory (Rubin & Wenzel, 1996), and the classical Stroop effect (MacLeod, 1991). These are only a few examples of replicated neuropsychological findings which are of importance for clinical practice. Moreover, replication

problems might differ depending on the clinical disorder one studies. A recent paper (Dumas-Mallet et al., 2016) included an elegant approach to test whether findings from original patient studies were replicated and how this differs across patient populations. In this study the authors went back to the original studies and tested whether the observed effects in these studies predicted the effect in corresponding meta-analyses. In psychiatry it seemed that the agreement between initial studies and corresponding meta-analyses was at chance level and for neurological and somatic diseases it was slightly better. Within psychiatry the highest replication rate was for cognitive and behavioral association studies (86%), and the lowest replication rate (6%) was observed for genetics. Due to the lack of available data a similar sub-analyses could not be run for neurological diseases. The specificity of larger studies was much better than for smaller initial studies, but no differences emerged when focusing on sensitivity. Hence, the reported power of the initial studies was not always the main problem when a finding failed to replicate.

Let's go back to clinical practice. 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. This issue of knowledge translation is a general issue within science, from clinical neuropsychology and biomedicine to sociology and economics, and is a research field on its own; see Graham and Tetroe (2007) for an overview.

5. Challenges

We see five major challenges in assessing experimental evidence within clinical neuro psychology.

The first challenge is that clinical tests can include a wide range of outcomes. For example, the Wisconsin Card Sorting Test 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 (e.g. total number of cards used, failure to maintain set, and perseverative errors) but also, just by chance, a difference on one specific measure (e.g. 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 issue that the difference between 'significant' and 'non-significant' is not

necessarily statistically significant (Gelman & Stern, 2006), a statistical fact of which researchers, journal editors, and policy-makers often seem unaware.

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. Exploratory research in itself is not problematic as long at is clearly labeled as such. Within clinical neuropsychological studies, one can often not randomize as we deal with pre-existing groups. This means that often interpretative additional analyses are run not to test the main hypothesis but to try to get a better understanding what is going on. This is of course fine as long as the authors do not report on these findings as if these were confirmatory. 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 (e.g. Gelman, Hill, & Yajima, 2012). Otherwise valuable data is being discarded. Displaying and integrating a grid of estimates complements the advice to display uncertainties as well as estimates for individual quantities. Likewise, 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. Therefore, we recommend using the raw data of the original studies when constructing a meta-analysis.

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. For example when there is sufficient time left in the research protocol one can add an experimental (e.g. a working memory task) from another study in order to replicate those findings while one's own main study has a completely different focus (e.g. facial recognition). As long as you clearly report which other tasks you included when discussing the findings of one of these studies and are transparent about the methods you used, two studies within one research protocol can be useful.

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 and clinical researchers to use these preregistration opportunities. Moreover, we recommend that researchers take, next to the APA guidelines, notice of existing guidelines for experiments (see e.g. http://www.consort-statement.org/) for observational studies (http://www.strobe-statement.org/) and to conduct meta-analyses and systematic reviews (http://prisma-statement.org/; see

for practical guidelines Gates & March, 2016). A special issue of *The Clinical Neuropsychologist* was dedicated to how to use and interpret existing research guidelines in a clinical neuropsychological context (Schoenberg, 2014). Another informative source is http://www.cebm.net/critical-appraisal from the Center for Evidence-based Medicine (CEBM) which provides worksheets available in different languages, to guide a systematic evaluation of clinical research papers.

But this does not resolve the question of when there is sufficient evidence to recommend adopting a specific treatment or to implement new tasks in an assessment protocol. Here we think it would make sense to perform a decision analysis, explicitly specifying costs, risks, and benefits, and constructing a decision tree in which gains and losses are weighted by the (estimated) probabilities of all outcomes. 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. With respect to the development of tasks within the field of clinical neuropsychological there is already extensive testing of validity and reliability as well as, for example, sensitivity and specificity of tasks before they are introduced to clinical practice. Moreover, specific handbooks such as Lezak et al. (2012) and Strauss et al. (2006) have been written to give clinicians an overview of the strengths and weaknesses of commonly used tasks in order to ensure that informed decisions can be made regarding what clinical tests to use. Moreover on the aforementioned CEBM website there is a worksheet available that enables a critical evaluation of diagnostic accuracy studies.

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 & Holtzman, 2013). However, findings from a meta-analysis cast doubt on this model (Schmand, Huizenga, & Van Gool, 2010). In the six years since its publication, this paper has not seemed to have had a serious impact on the field. When we can believe a research finding, how much replication is needed to counterbalance methodological criticisms and unsuccessful replications, and when is evidence sufficient and sound enough to transfer knowledge to the clinic? From one direction, research protocols should be made more realistic so that research findings have more direct clinical application; from the other direction, the principles of quality control can be applied to clinical practice, with data routinely recorded and analyzed observationally. This is already becoming more and more common in various clinical institutions.

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 difference, 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? Researchers can, for example report, next to the group score, how much each individual participant differs from a normative group. A well-known method is the regression-based summary score to assess neuropsychological differences or change (see e.g. Cysique et al., 2011; Duff, 2012; McSweeny, Naugle, Chelune, & Luders, 1993). In this case one compares one test score with a normative score, taking, for example, sex and education

level into account. Another well-known statistical method to determine whether there is an actual change is by calculating a reliable change index (e.g. Duff, 2012). However, neuropsychologists are often interested in learning about changes within a cognitive profile. Currently, 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 (multivariate normative comparisons, see Huizenga, Agelink van Rentergem, Grasman, Muslimovic, & Schmand, 2016) are already implemented so the users can use these statistical techniques without knowing all the details of these techniques. This can also be used by researchers to show how many patients actually show changes in their cognitive profile.

6. 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 positive effects for children with ADHD (Klingberg, Forssberg, & 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 vs. 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 (e.g. Buschkuehl & Jaeggi, 2010; Chacko et al., 2013; Evans et al., 2013; Melby-Lervag & Hulme, 2013; Morrison & Chein, 2011; Rabipour & Raz, 2012; Rapport, Orban, Kofler, & Friedman, 2013; Shipstead, Hicks, & Engle, 2012; Spencer-Smith & 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 (2015) 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 and correcting for the publication bias (Dovis, Van der Oord, Huizenga, Wiers, & Prins, 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 (see also Dumas-Mallet et al., 2016) 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 extensive reviews regarding a specific topic.

7. 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. The examples in the introduction demonstrate the ways in which scientifically and statistically dubious claims can make it through the peer-review process at respected journals (see Smith, 2006) and be promoted in leading media outlets (see e.g. Yong, 2012; Gelman, 2015, 2016b). Even for silly published claims on topics such as power pose, extra-sensory perception, or embodied cognition that might not affect policy or change our own behaviors, we do need to take this problem seriously as it reduces the trust in the coherence of science, among working researchers as well as the general public.

Awareness of the statistical crisis in psychology research is not new (see e.g. Meehl, 1990, and he had been writing about this for decades by then), but it is now at the center of attention. In recent years we have come to realize that many seemingly solid results in various research fields 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 & 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. Design analysis is not an alternative to replication but rather is an improved attempt to understand the information in available data so as to have a better sense of reproducibility.

Recent papers by Ioannidis (2005), Simmons et al. (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 translate

across disciplines to raise additional awareness about good research practices. Even though within the field of clinical neuropsychology there has been a lot of attention for many of the aforementioned issues over the years, clinicians and researchers alike need reminders to ensure such good research practice.

Some of the suggested solutions are already incorporated in guidelines of the American Psychological Association, the CEBM website, and of specific journals, but much more needs to be done. 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 must continue with careful analysis of historical and new data, continue replication of existing studies (and publication of the results of these replication attempts), and have replication in mind when designing new studies. We need to move away from the idea of statistical significance as a demonstration of effectiveness and instead think of each new study as part of a learning process.

We hope that, years from now, no one will agree with Tulving and Madigan (1970), who, after reviewing the literature on memory, declared that for two third of the papers 'the primary function these papers serve is giving something to do to people who count papers instead of reading them.' So read publications critically and make sure that your own methods sections are detailed enough to ensure that others can also critically evaluate your findings.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially supported by the National Science Foundation; Institute of Education Sciences; Office of Naval Research; and the Netherlands Society for Scientific Research.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*, 452–454.

Bargh, J. A. (2012). Priming effects replicate just fine, thanks. The Natural Unconscious blog, *Psychology Today*, May 11. Retrieved from https://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A), 23*, 399–406.

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly, 140*, 266–272.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*, 1363–1368.

Chacko, A., Feirsen, N., Bedard, A. C., Marks, D., Uderman, J. Z., & Chimiklis, A. (2013). Cogmed working memory training for youth with ADHD: A closer examination of efficacy utilizing evidence-based criteria. *Journal of Clinical Child and Adolescent Psychology, 42*, 769–783.

Chen, P., & Goedert, K. M. (2012). Clock drawing in spatial neglect: A comprehensive analysis of clock perimeter, placement, and accuracy. *Journal of Neuropsychology, 6*, 270–289.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180.

Cysique, L. A., Franklin, D., Jr., Abramson, I., Ellis, R. J., Letendre, S., Collier, A., … Simpson, D. (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology, 33*, 505–522.

Dovis, S., Van der Oord, S., Huizenga, H. M., Wiers, R. W., & Prins, P. J. (2015). Prevalence and diagnostic validity of motivational impairments and deficits in visuospatial short-term memory and working memory in ADHD subtypes. *European Child and Adolescent Psychiatry, 24*, 575–590.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*, e29081.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology, 27*, 248–261.

Dumas-Mallet, E., Button, K., Boraud, T., Munafo, M., & Gonon, F. (2016). Replication validity of initial association studies: A comparison between psychiatry, neurology and four somatic diseases. *PLoS ONE, 11*, e0158064.

Evans, S. W., Brady, C. E., Harrison, J. R., Bunford, N., Kern, L., State, T., & Andrews, C. (2013). Measuring ADHD and ODD symptoms and impairment using high school teachers' ratings. *Journal of Clinical Child and Adolescent Psychology, 42*, 197–207.

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

Gates, N. J., & March, E. G. (2016). A neuropsychologist's guide to undertaking a systematic review for publication: Making the most of PRISMA guidelines. *Neuropsychology Review, 26*, 1–12.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*, 632–643.

Gelman, A. (2016a, January 26). The time-reversal heuristic – A new way to think about a published finding that is followed up by a large, preregistered replication (in context of Amy Cuddy's claims about power pose). *Statistical Modeling, Causal Inference, and Social Science Blog*. Retrieved from http://andrewgelman.com/2016/01/26/more-power-posing/

Gelman, A. (2016b, February 1). When does peer review make no damn sense? *Statistical Modeling, Causal Inference, and Social Science Blog*. Retrieved from http://andrewgelman.com/2016/02/01/peer-review-make-no-damn-sense/

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*, 189–211.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460–465.

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

Graham, I. D., & Tetroe, J. (2007). Some theoretical underpinnings of knowledge translation. *Academic Emergency Medicine, 14*, 936–941.

Hageman, W. J., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy, 37*, 1169–1193.

Hugdahl, K. (2011). Fifty years of dichotic listening research – Still going and going and …. *Brain and Cognition, 76*, 211–213.

Huizenga, H. M., Agelink van Rentergem, J. A., Grasman, R. P., Muslimovic, D., & Schmand, B. (2016). Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives. *Journal of Clinical and Experimental Neuropsychology, 38*, 611–629.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.
Ioannidis, J. P. A., & Panagiotou, O. A. (2011). Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA, 305*, 2200–2210.
Jack, C. R., & Holtzman, D. M. (2013). Biomarker modeling of Alzheimer's disease. *Neuron, 80*, 1347–1358.
Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., … Westerberg, H. (2005). Computerized training of working memory in children with ADHD-A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*, 177–186.
Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A), 24*, 781–791.
Lezak, M. D., Howieson, D., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.
McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Luders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist, 7*, 300–312.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.
Melby-Lervag, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270–291.
Millis, S. (2003). Statistical practices: The seven deadly sins. *Child Neuropsychology (Neuropsychology, Development and Cognition: Section C), 9*, 221–233.
Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin and Review, 18*, 46–60.
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 1–8.
Rabipour, S., & Raz, A. (2012). Training the brain: Fact and fad in cognitive and behavioral remediation. *Brain and Cognition, 79*, 159–179.
Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*, 653–656.
Rapport, M. D., Orban, S. A., Kofler, M. J., & Friedman, L. M. (2013). Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes. *Clinical Psychology Review, 33*, 1237–1252.
Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734–760.
Schatz, P., Jay, K. A., McComb, J., & McLaughlin, J. R. (2005). Misuse of statistical tests in archives of clinical neuropsychology publications. *Archives of Clinical Neuropsychology, 20*, 1053–1059.
Schmand, B., Huizenga, H. M., & Van Gool, W. A. (2010). Meta-analysis of CSF and MRI biomarkers for detecting preclinical Alzheimer's disease. *Psychological Medicine, 40*, 135–145.
Schoenberg, M. R. (2014). Introduction to the special issue on improving neuropsychological research through use of reporting guidelines. *The Clinical Neuropsychologist, 28*, 549–555.
Shabbir, S. H., & Sanders, A. E. (2014). Clinical significance in dementia research: A review of the literature. *American Journal of Alzheimer's Disease and Other Dementias, 29*, 494–497.
Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition, 1*, 185–193.
Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Simmons, J., & Simonsohn, U. (2015, May 8). Power posing: Reassessing the evidence behind the most popular TED talk. *Data Colada blog*. Retrieved from http://datacolada.org/37

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*, 178–182.

Spencer-Smith, M., & Klingberg, T. (2015). Benefits of a working memory training program for inattention in daily life: A systematic review and meta-analysis. *PLoS ONE, 10*, e0119522.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society.

Tulving, E., & Madigan, S. A. (1970). Memory and verbal learning. *Annual Review of Psychology, 21*, 437–484.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. J. (2015). A skeptical eye on psi. In E. May & S. Marwaha (Eds.), *Extrasensory perception: Support, skepticism, and science* (pp. 153–176). ABC-CLIO.

Yong, E. (2012). In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature, 485*, 298–300.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain, 113*, 1721–1777.

Zihl, J., & Heywood, C. A. (2016). The contribution of single case studies to the neuroscience of vision. *PsyCh Journal, 5*, 5–17.