Psychological Science's Replicability Crisis and What it Means for Science in the Courtroom

Jason M. Chin

University of Toronto

March, 2014

In Press, Journal of Psychology, Public Policy, and Law

Introduction and Overview

Science – psychological science especially – is undergoing a process of massive change. Spurred by startling revelations about the dependability of its findings (i.e., the "replicability crisis"; see Pashler & Wagenmakers, 2012 for a review),[1] generally accepted practices are being reformed at a pace and ferocity with scant precedent. Throughout this process longstanding recommendations (Cohen, 1994) and norms (American Psychological Association, 2009; Wilkinson Task Force, 1999) have become binding as academic journals – the very outlet for scientific work – have changed their review process and requirements for publication (e.g., *Nature*, *Science*, *Psychological Science*). In tandem with these responses, well-funded organizations have sprouted up providing further research and infrastructure to scaffold these changes (e.g., *The Open Science Framework*). While the swift reaction to the crisis is laudable, the fact that evidently substandard generally accepted practices persisted for so long poses important questions concerning how law should treat scientific evidence, and how it should react to the changes occurring in science.

In evaluating scientific evidence, courts rely on both testimony from qualified experts in a field (Federal Rules of Evidence, Rule 702, 2011) and after *Daubert v. Merrell Dow* ("*Daubert*"; 1993), the judge's own evaluation of the validity of this evidence. In fact, *Daubert* and its two companion cases (*General Electric Co. v Joiner*, 1997; "*Joiner*"; *Kumho Tire Co. v Carmichael*, 1999; "*Kumho*"; i.e., the "the *Daubert* trilogy") are the leading U.S. Supreme Court authority on the admissibility of scientific evidence and are binding on federal courts.[2] *Daubert*'s innovation is the requirement that judges play a gatekeeper role, not just determining

[1] The term "crisis", while certainly strong, tracks language used in the field and is thus useful in succinctly conveying meaning in the present Essay. But it must be noted that, perhaps due to science taking this "crisis" so seriously, a great deal of progress has been made and this crisis may be abating.

[2] But as shall be described later in the Essay, several states have refrained from adopting the *Daubert* framework.

whether a scientific finding or method has reached a sufficient consensus among scientists, but evaluating the validity of the science itself. Posing both important practical considerations, and questions fundamental to the philosophy of science, *Daubert* has attracted a great deal of academic interest (e.g., see Beecher-Monas, 1998: Chesebro, 1994; Faigman & Monahan, 2009; Schwartz, 1997).

The replicability crisis, however, exposes a fundamental concern with *Daubert*. In particular, it suggests that generally accepted practices also affect the *way* in which research is conducted, thus complicating judicial analysis of scientific findings. As shall be described below, the most important prescription this Essay has to offer is that the legal treatment of scientific evidence track new measures developed in science to improve its dependability. This is especially important because many of the concerns raised during the replicability crisis have been raised before over the past several decades, with little substantive change to show for it. Moreover, while the scientific stakes of allowing unreliable findings to permeate bodies of research are high, the stakes are just as high within law where, outside of limited appellate review, there is less opportunity for self-correction. Law must take measures to ensure that the science that reaches decision makers is a result of best practices, and not merely those that are generally accepted.

To address the above goal, this Essay follows a simple structure. After a brief review of the law and the replicability crisis, I analyze the *Daubert* factors in light of the causes and lessons of the replicability crisis. This analysis suggests that the deficits that gave way to the replicability crisis have also exposed weaknesses in the *Daubert* framework. These revelations have implications for both evidence that enters the courtroom through experts appointed by parties and through a judge's own fact-finding endeavors. Practically speaking, if judges,

lawyers and jurors are not wary of replicability issues, bad science will almost certainly impact the legal process. I end with a few tentative prescriptions for how the law might evolve in light of the replicability crisis.

Scientific Evidence in the Courtroom: From *Frye* to *Daubert*

Daubert has been described as the "foundational opinion" in the field of scientific evidence and the law (Cheng & Yoon, 2005, p. 471). The *Daubert* trilogy represents the juristic standard for the admission of scientific evidence in federal and many state courts in the U.S. (Berstein & Jackson, 2004), and is binding in Canada as well (see Goudge, 2008). In *Daubert*, the Supreme Court modified the standard for the admission of scientific evidence formerly found in *Frye v United States* ("*Frye*"; 1923). *Frye* required an inquiry into whether a scientific finding or practice had reached general acceptance in the scientific community, asking if it was "sufficiently established to have gained general acceptance into the field in which it belongs" (*Frye*, 1923, p. 2796).

The *Daubert* Analysis

The Court in *Daubert* overhauled *Frye*, requiring judges look not only at general acceptance, but into the reliability of the scientific evidence itself. *Daubert* offered four factors potentially useful in assessing the scientific validity of proffered evidence: (1) the testability of the scientific knowledge; (2) peer review and publication of the evidence; (3) the error rate associated with the evidence, and; (4) the general acceptance of the finding or practice (*Daubert*, 1994, p. 2797). In the second of the *Daubert* Trilogy, the Court in *Joiner* confirmed that in a *Daubert* analysis, a court is free to assess the expert's opinion, the data supporting it, and the connection between the two (*Joiner,*1997, p. 143). The Court also held that a trial judge's

application of *Daubert* should be reviewed on the "abuse of discretion" standard (*Joiner*, 1997, p. 142). This high standard makes it difficult for appellate courts to overturn a trial judge's decision to admit expert testimony, placing increased importance on this decision at first instance. The Supreme Court then elaborated in *Kumho*, stating that the *Daubert* factors applied to technical expertise as well, and that they should be considered a flexible and non-exhaustive set of considerations (*Kumho*, 1999, p. 1179).

Adoption of *Daubert* remains a live issue at the state level. Although many state courts have adopted *Daubert*, several have expressly rejected it for *Frye* (e.g., California, Michigan and New York). And among those state courts that have adopted *Daubert*, several have not adopted the full trilogy. For instance, many states have adopted *Daubert* and *Joiner*, but not *Kumho* (Bernstein & Jackson, 2004). As facts arise that engage various elements of the trilogy, it can be expected that its adoption will continue to be fought at the state level because of the general consensus that *Daubert* represents a stricter standard (Sanders, Diamond, & Vidmar, 2002).

Indeed, *Daubert* adds three factors to bolster general acceptance, which was held over from *Frye*. The peer review and publication factor was added because "…submission to the scrutiny of the scientific community is a component of 'good science,' in part because it increases the likelihood that substantive flaws in the methodology will be detected" (*Daubert*, 1994, p. 2797). The testability factor has been interpreted to mean both whether a question can be scientifically falsified, and whether the evidence has been subjected to proper research methods, including the internal and external validity of the studies (Beecher-Monas, 1998). Finally, the error rate factor can encompass a variety of statistical means of supporting a scientific technique or finding (Beecher-Monas, 1998, p. 71; Faigman & Monahan, 2009, p. 13).

Again, *Daubert* is widely considered to represent a move from deference to general acceptance of scientific findings, to judicial review of the validity of the science itself.

Frye & *Daubert*: The Role of the Trial Judge

Spanning an array of issues applicable to the sciences, law, and the philosophy of science, there is no shortage of academic treatment of *Daubert* (e.g., see Beecher-Monas, 1998; Faigman, 1999; Cheng & Yoon, 2005). Indeed, in asking judges to evaluate the quality of science, *Daubert* inspires questions regarding the very nature of scientific inquiry. These questions are generally beyond the scope of this Essay, which is a more discrete inquiry into the failure of specific scientific safeguards and practices, how such failures can lead to findings that are difficult to replicate, and how this should inform law's treatment of scientific evidence. This conceptualization of the scientific method is perhaps best encapsulated by a view that discriminates between good and bad science as differences in degree, rather than kind (Lilienfeld & Landfield, 2008; Lilienfeld, Lynn, & Lohr, 2003). In fact, following from this view, Lilienfeld and Landfield's (2008) *indicia* of pseudoscience (as compared to science) have interesting reflections in the sources of the replicability crisis. For instance, just as pseudoscience often seeks to evade peer review (Lilienfeld, Lynn, & Lohr, 2003, p. 6), we shall see that impoverished peer review helped contribute to the replicability crisis. With this in mind, I now describe how *Daubert*'s innovation was treated by academics.

The immediate reaction within academia to *Daubert* was divided over whether it would prove a more discerning standard than its predecessor (see Chesebro, 1994 for a review). For example, in a series of toxic tort cases prior to *Daubert*, courts had applied *Frye* quite rigidly, excluding evidence deemed too novel to have reached general acceptance (*Christophersen v Allied-Signal Corp*, 1991; *Sterling v Velsicol Chem. Corp.*, 1988). This prompted academics to

worry that the additional factors in *Daubert* added flexibility and would ultimately soften the standard (Chesebro, 1994).

Other commentators greeted *Daubert* warmly. Beecher-Monas (1998), for example, suggested *Frye* was too easily abused as it was not difficult to assemble enough experts to demonstrate general acceptance. She also found that several courts initially applied *Daubert* in a scientifically sophisticated manner, going beyond mere acceptance to cogently examine the methods and data analysis of the proffered evidence. In support of her argument, Beecher-Monas noted that a *Daubert* analysis of evidence regarding the social psychology of coerced confessions was well-applied (*United States v Hall*, 1997). Still, some researchers have argued that the impact of a jurisdiction's decision to adopt *Daubert* over *Frye* is limited by the scientific fluency of legal actors.

Research regarding the capability of judges, lawyers and jurors of understanding and applying the scientific concepts that underlie the *Daubert* test (see McAuliff & Groscup, 2009 for a review) has yielded less than promising results. These results converge with well-established findings indicating human cognition is susceptible to several biases that may interfere with reasoning about science (e.g., see Kahneman & Tversky, 1973; Likwornik, 2012). With regard to judges, Gatowski (2001), performed a survey study finding many showed a lack of scientific understanding of the error rate and testability concepts, but moderate to strong understanding of peer review and general acceptance. Judges seemed to also prefer th traditional *Frye* analysis with 42% of judges choosing general acceptance as the most important factor (Gatowski, 2001, p. 448).

Using experimental methods, Kovera and McAuliff (2000) measured the admissibility decisions of Florida circuit court judges in response to expert testimony that the experimenters had manipulated to reflect different aspects of *Daubert*. Depending on condition, the expert testimony was purportedly either published in a peer reviewed journal or not, and contained varying levels of internal validity (e.g., a missing control group). The results were once again discouraging in terms of the judges' use of components of the *Daubert* framework. Lack of internal validity and publishing appeared to have little or no effect on the judge's decision to admit the evidence. And moreover, less than 15% appeared to notice the flaws with regard to internal validity when they were present.

Similar research has been performed assessing the way lawyers (described in McAuliff & Groscup, 2009) and jurors (Kovera, McAuliff, & Hebert, 1999; McAuliff & Kovera, & Nunez, 2009) assess scientific evidence with results in line with the findings on judges. With regard to lawyers, Kovera and McAuliff (described in McAuliff & Groscup, 2009) utilized the same materials and methods as with the above study on judges, finding converging evidence that lawyers were unmoved by the internal validity and peer review factors when deciding to file a motion to exclude expert testimony. They report that lawyers did, however, appear attentive to the publishing factor, rating such findings as more reliable.

Jury-eligible individuals asked to evaluate scientific evidence were mostly insensitive to internal validity and peer review, although they did seem to give less weight to research lacking a control condition (McAuliff, Kovera, & Nunez, 2009). In a different study, Kovera, McAuliff and Hebert (1999) found that jury-eligible university students viewing a simulated trial partially based their decision making on the publication status and ecological validity (i.e., the fit between

the participants in the study and the parties to the dispute) of the scientific evidence when evaluating it.

The above research is pessimistic regarding the scientific fluency of judges, lawyers and jurors. Still, this is a field in its infancy and researchers readily admit that it is limited in both its quantum and the methodologies used (McAuliff & Groscup, 2009, p. 31). For example, the data from the McAuliff and Kovera research program were largely culled from the context of litigation arising from workplace disputes. Moreover, some research suggests that there are ways to frame statistical concepts in a manner that is more easily digested by non-scientists (Cosmides & Tooby, 1996). After a discussion of the replicability crisis, I will review specific avenues that present hope in bridging the divide between law and science.

In addition to research focused on scientific fluency, other work has examined trends in state and federal courts, because as established above, several states have refrained from adopting *Daubert*. This research has generally found little to no effect of *Daubert* on admission decisions. For example, in the context of criminal cases, Groscup and colleagues (2002) found state adoption of *Daubert* did not affect admission rates. Moreover, Cheng and Yoon (2005) report data suggesting that state court adoption *Daubert* over *Frye* matters little because it does not seem to impact litigant behavior.

As I shall explain below, concerns with the ability of judges to apply *Daubert* effectively may be corrected by a multitude of measures, such as court appointed technical advisors and improved judicial training on scientific topics. The replicability crisis, however, presents a novel and distinctive type of threat. In particular, it suggests that several generally accepted practices have long lagged behind best practices. This phenomenon has subtly affected – to its detriment –

how science is performed and communicated. Fortunately, equipped with an awareness of replicability issues, initiatives designed to improve *Daubert*'s application have a high likelihood of success. Prior to elaborating on this notion, a brief primer on the replicability crisis is useful.

Replicability Crisis Overview

Concerns with flaws in the research and publication process in the sciences are far from new, with several calls for change having been made over the years (e.g., American Psychological Association, 2009; Cohen, 1994; Wilkinson Task Force, 1999). The present crisis, however, is notable both for the sheer amount of research it has inspired plumbing its depths and causes, and for the tangible change it has inspired. Both have significant implications for the law's treatment of scientific evidence. While I will focus on the replicability crisis as it was felt in psychology, it must be noted that it is certainly not limited to psychology (e.g., see Ioannidis, 2005). The focus on psychology is partially one of convenience – psychological scientists have been especially concerned with the issue (Yong, 2012) and have produced a great deal of empirical work examining the crisis (see Pashler & Wagenmakers, 2012 for a review). Furthermore, most of the discussion will focus on the processes underlying the replicability crisis as projects have only begun to identify unreplicable findings with sufficient certainty (but see a promising start by Klein et al, 2013).

It is difficult to say exactly when the replicability crisis began. Some writers (Funder et al, 2013) trace it to an article by Ioannidis (2005) titled "Why Most Published Research Findings are False." In this article Ioannidis provided quantitative models indicating a high likelihood that many well accepted studies are in fact false. Following the Ioannidis paper, several high-profile discoveries converged to cast a great deal of doubt on the efficacy of research and publication

standards. These doubts culminated in the past two years with a special edition of a leading journal devoted to the crisis (Pashler & Wagenmakers, 2012) and new non-binding guidelines from the *Society for Personality and Social Psychology* ("*SPSP*"; Funder et al, 2013). Journals have followed suit with changes in the editorial policy of such high-profile academic outlets as *Nature* (Editorial, 2013), *Science* (McNutt, 2014), and *Psychological Science* (Eich, 2013).

In 2009, work by Vul, Harris, Winkielman and Pashler found that correlations in a burgeoning area of research that uses fMRI to study social psychological processes were higher than would statistically be expected given how the measurements were taken and analyses performed. Briefly, the source of the problem appears to be that experimenters naively obtained measurements in a non-independent manner. In particular, several studies in Vul and colleagues' review based their analysis on fMRI measurements that had been selected *because* they met a threshold determined by the experimenters. In other words, these measurements were chosen due to their relation to a variable of interest rather than in the independent manner that their statistical analyses assume. The results were, therefore, biased in favor of finding a statistically significant result when one did not exist. Following the controversy caused by the Vul and colleagues (2009) article, a number of issues with the scientific process were highlighted in the media.

Notably, in 2010, the *New Yorker* published an article titled "The Truth Wears Off." It presented a description of the "decline effect" and its impact on several lines of research (Lehrer, 2010). The decline effect, as the name suggests, is a phenomenon in which effect sizes (i.e., the magnitude of the observed difference between an experimental condition and a control condition) of scientific findings will sometimes decrease over time. This article focused on the declining size of an eyewitness memory phenomenon called "verbal overshadowing." Verbal

overshadowing is the finding that when one verbally describes a face, he or she tends to be less accurate in correctly identifying that face in a lineup afterwards, relative to someone who did not describe it (Chin & Schooler, 2008). Lehrer noted that between 1990 and 1995, the size of the verbal overshadowing effect dropped by 30%, and then another 30% in the following years. This account accords with a meta-analysis (i.e., a statistical review of published and unpublished data) performed by Meissner and Brigham (2001), which saw the effect downgraded from a medium to small effect size. During this time other media outlets reported additional anecdotal difficulties with replication (e.g., Bower, 2012; Yong, 2012).

Media attention coincided with the discovery of scientific fraud perpetrated by a well-respected social and cognitive psychologist. In August 2011, an investigation determined that at least 34 of Diedrick Stapel's published studies were based on wholesale data fabrication (Stroebe, Postmes & Spears, 2012, p. 670). The scandal did little to improve psychology's public image. For example, Benedict Carey writing for the *New York Times* commented that the Stapel incident "…exposes deep flaws in the way science is done in a field…" (Carey, 2011). Media attention stemming from the replicability still continues (e.g., see Johnson in the *New York Times*, 2014).

An additional high-profile case of reproducibility issues occurred when Daryl Bem published an article in social psychology's flagship journal presenting nine experiments with over 1,000 total subjects supporting the existence of extra-sensory perception ("ESP"; 2011, p.407). Bem temporally reversed the order of several classic psychological studies, finding that experimental manipulations that would occur later in the study impacted present behavior. These highly implausible findings attracted a great deal of attention both within and without the

scientific community. Many took this apparently strong evidence for ESP as a sign that current standards for reporting data were insufficient.

Wagenmakers and colleagues (2011) responded to the Bem paper, suggesting several reasons Bem may have achieved false positives in eight of the nine studies (2011, p.426). These researchers first noted that many of the analyses in the ESP paper were exploratory, rather than confirmatory. Exploratory research derives not from *a priori* hypotheses, but from a look at the data itself. Wagenmakers and colleagues point out that exploratory practices increase the chances of false positives, and should only be used if they are followed with confirmatory research. They found several indications of exploratory research in the Bem paper (Wagenmakers, 2011). Subsequent attempts at reproduction failed to replicate Bem's findings (Galak, 2012; Ritschie, Wiseman, & French, 2012; Robinson, 2011).

After these highly visible (but largely anecdotal) examples of scientific dysfunction, psychology responded with a freshet of systematic work examining how such failures could be possible, and how prevalent they are (e.g., Pashler & Wagenmakers, 2012). Correspondingly, journals and governing bodies took steps to address the crisis (e.g., Funder et al, 2013). In the subsequent section I will review the sources of the replicability crisis with the *Daubert* factors as an armature for my analysis.

Sources of The Replicability Crisis, Through the Lens of *Daubert*

The issues that contributed to the replicability crisis in science pose problems for *Daubert*. While the *Daubert* does provide a coherent framework for evaluating a scientific finding, I will demonstrate that the *Daubert* Court made assumptions about the scientific practices underlying that framework that have, in some cases, been violated. In particular, generally accepted practices for testing, error reduction and peer review have at times not tracked

best practices. In this section, I review the *Daubert* factors in light the lessons learned from the replicability crisis. Most attention will be paid to testability and error rates, as those represent *Daubert*'s primary innovations.

General Acceptance

A chief concern among critics of *Frye*'s general acceptance standard was that by deferring the evaluation of scientific evidence to expert consensus, it represented little more than an exercise in nose-counting (Faigman & Monahan, 2009). In other words, general acceptance opens up the possibility that a technique or result, although generally accepted in a field, had not been rigorously subjected to the scientific method. Lessons from the replicability crisis bolster these qualms with general acceptance.

The replicability crisis was made possible by the failure of generally accepted practices to track best practices. Indeed, the concerns brought up during the replicability crisis can be traced back for decades. For example, as shall be discussed in greater depth below, use of overly simplistic and misleading statistical procedures were decried consistently by experts (e.g., Cohen, 1994). These concerns were eventually echoed by an *American Psychological Association* Task Force (Wilkinson Task Force, 1999) with explicit recommendations for overhauling statistics in psychological research. The very same recommendations are being made once again in reaction to the replicability crisis (Funder et al, 2013).

The pattern of recommended change and subsequent disregard for these prescriptions illustrates that generally accepted practices do not always reflect best practices. A useful parallel may be found in transnational law where there is often a lag between non-binding norms and their adoption by binding and enforceable agreements (e.g., the U. N. *Convention on*

Corruption's adoption into the World Trade Organization *General Procurement Agreement*). Commentators have suggested that the problem in science may stem from a publication system that prioritizes style over substance (Bakker, van Dijk & Wicherts, 2012). Regardless, it is a story that reaffirms the need for a legal standard that goes beyond general acceptance. Unfortunately, as argued below, the replicability crisis demonstrates that flaws in generally accepted practices can, in some circumstances, contaminate the way science is performed and disseminated (i.e., the other three *Daubert* factors). Therefore, even a wary judge or scientist will have trouble determining which findings are dependable and which are not.

Publication and Peer Review

To quote *Daubert*, the rationale behind the publication and peer review factor is: "…submission to the scrutiny of the scientific community is a component of 'good science,' in part because it increases the likelihood that substantive flaws in the methodology will be detected." (*Daubert*, 1994, p. 2797). In other words, this factor seeks to screen scientific evidence by considering whether the science has been disseminated and vetted. The replicability crisis demonstrates that peer review and publication is problematic because it is subject to generally accepted practices that have long lagged behind best practices. For example, peer review has traditionally not required that raw data be subjected to the peer review process, a system that allows researchers a great deal of latitude in how they analyze and present their data. What follows describes some of these issues with peer review and publication that contributed to the replicability crisis.

Lack of Access to Raw Data

Researchers believe that lack of access to raw data played a part in the replicability crisis (Funder et al, 2013). In other words, statistical analyses could not be vetted by methodologists because the raw data was not available. Contrary to what many would expect, journals have not always required that raw data be produced along with the report associated with it. Methodologists have long cried for such transparency, noting that the cases in which such a requirement would be unwieldy (e.g., proprietary data, privacy issues) are few and far between (Wicherts & Bakker, 2012).

The lack of such a requirement would not be as detrimental if authors were forthcoming with their data when contacted. However, recent research (Wicherts, Bakker & Molenaar, 2011) found that less than 50% of authors shared their data, despite multiple requests. Moreover, it was the authors with statistically weaker findings that declined to share. Simonsohn (in press) recently noted that access to data is extremely useful in ruling out benign explanations when results seem suspicious. Provision of raw data would help scientists do precisely what the Court in *Daubert* thought peer review and publishing would achieve: assess scientific validity. Moreover, access to raw data would assist expert witnesses in creating a veridical picture of a scientific field to present to legal decision makers, rather than relying on the analyses performed by the authors of the work.

Publication Bias

Another likely source of the replicability crisis is publication bias (i.e., a prejudice that can form in a body of knowledge when a certain type of result is preferred over another; Dickersin, 1990). One variety of publication bias is a bias towards publishing positive over negative studies. Indeed, it is well accepted that negative studies (i.e., those showing no statistically significant result) are often discriminated against, with positive studies (i.e., those

showing an effect) predominating the literature (Fanelli, 2012; Sterling, 1959). In fact, literature reviews often must provide some kind of *ad hoc* correction for this phenomenon that is known as the "file-drawer effect" to indicate that negative studies are never seen outside the confines of the laboratory that produced them (Rosenthal, 1979). Despite such corrections, the file-drawer effect poses a serious problem for accurately assessing the true nature of a scientific finding. In other words, the peer review and publication process, by being biased towards positive results, allows for overestimation of the strength and robustness of a finding or method to reach general acceptance.

Empirical evidence supports the existence of a publication bias in scientific bodies of research. For example, Sterling (1959) examined four psychological journals during the course of a year and found that articles reporting data that failed to reject the null hypothesis comprised only 2.72% of the articles published. Several findings in other fields suggest that this is not an uncommon result (e.g., see Dickersin, 1990a; Coursol &Wagner, 1986). Research tracking the percentage of positive results across all fields finds the proportion has hovered around 90% in recent years (Fanelli, 2012). Experimental evidence confirms such reports. For instance, Mahoney (1977) randomly assigned journal referees to read and evaluate articles that were subtly tweaked to contain positive or null results. He found that referees evaluating work with negative results not only deemed the overall paper to be less worthy of publication, but also rated the methods as less convincing – despite the fact that the methods sections of the papers were identical. These policies have a clear effect on authors. Greenwald (1975), in surveying a sample of psychological researchers, found that 41% reported they would seek to publish a positive finding, whereas only 5% would seek to publish a null result.

Ironically, by placing weight on peer review and publication, *Daubert* may indeed distract judges from bodies of unpublished work that contain valuable data (e.g., negative findings, successful and failed replications). In addition, recall that one of the few facets of the scientific process jurors are attentive to is the publication status of research (Kovera, McAuliff, & Hebert, 1999). Even if unpublished research passes the judicial gatekeeper, there is worry that it will be undervalued by jurors. The following section suggests other ways in which publication and peer review may be misleading to judicial actors.

Valuing Novelty over Rigor

A focus on novel and exciting results and corresponding lack of attention to methodology in the peer review and publication process has been implicated as a source of the replicability crisis (Giner-Sorolla, 2012). The changes *Nature* is making to its review process illustrates the problem. In fact, *Nature* has explicitly admitted the journal's role in the replicability crisis:

> "The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly." (Editorial, 2013, p .398)

The journal *Psychological Science* has made similar admissions. In response to the replicability crisis, this journal recently revised its word limits, exempting methods and results sections from the limit. This monumental change was followed with the commentary:

> "The intent here is to allow authors to provide a clear, complete, self-contained description for their studies, which **cannot be done** with restrictions on Method and Results." (Eich, 2013, p. 1; emphasis is my own)

Similar changes can be found in new guidelines being produced by professional organizations. For instance, SPSP's Task Force on Publication and Research Practices recommendation four reads:

> "Include in an appendix the verbatim wording (translated if necessary) of all the independent and dependent variable instructions, manipulation and measures." (Funder et al, 2013, p. 6)

These explicit and implicit admissions that previous practices were giving insufficient attention to methodology are troubling from a legal standpoint. In particular, according to *Daubert*, a trial judge's gatekeeper role focuses "…solely on the principles and methodology; not on the conclusions they generate" (1993, p. 1797). One wonders how well the publication and peer review factor can signal that a methodology has been reviewed if that methodology is not made fully available to reviewing scientists. And further, how should a gatekeeper assess methodology if only the publishing scientist is privy to that information?

Complicating the matter are initiatives designed to both make science more readily accessible, but lack a substantive peer review component. One example is the relatively recent phenomenon of open access journals. These academic outlets are unlike standard academic journals in that they are not situated behind pay-walls and are open to the public to access and, ostensibly, to critique. They rest on an economic model by which authors pay the costs of publication. These journals should have the effect of allowing traditionally non-publishable data to see the light of day. The downside is that the level of peer-review at these journals seems to be substandard in many cases. For instance, *Science* recently published a finding indicating that over 50% of open access journals were willing to publish a "bait" study they had submitted with

serious errors of design and data interpretation (Bohannon, 2013). Open access journals therefore may assist courts insofar as they make more data available. But at this point, deference to the review process at these journals may be problematic.[3]

It has been noted before that peer-review and publication does not necessarily lead to research that is beyond reproach (e.g., Faigman & Monahan, 2009; Jasanoff, 1996). The replicability crisis demonstrates that this issue may be more serious than previously thought. For one, despite the *Daubert* Court's apparent belief that methodology would be well vetted (i.e., a close review of the design, variables studied and how the data was analyzed), methodological review has been lacking. This phenomenon has been documented in the past, with critics lamenting a shift from a more fulsome methods-orientated science reporting to the current results-driven model (Dickersin, 1990). These are precisely the issues *Nature* and *Psychological Science* recently grappled with, deciding that its existing review process did not look deeply enough into methodology. While these changes are welcome, the tacit admission that previous guidelines were lacking is troubling for *Daubert*'s peer review and publication factor. Going forward it will be important that these new practices be imported into how the legal field treats scientific evidence.

Testability and Error Rates

The sources of the replicability crisis also indicate that the testability and error rate factors are of less use than originally thought. In theory, analyzing whether and how a finding or method has been tested should allow judges to readily weed out results that are not sufficiently reliable. These concepts were presented in a somewhat nebulous manner in the *Daubert* Trilogy (Faigman & Monahan, 2009). In addition, there is also a worry that such a task is beyond the

[3] There are, however, open access journals (e.g., PLOS ONE, cited numerous times in this Essay) that routinely engage in demonstrably high levels of peer review.

scientific competence of judges (Gatowski, 2001). While this concern may be rectified by training and the use of court-appointed experts in complicated cases, the replicability crisis presents a more fundamental concern. In particular it poses the problem that the testing and error rate standards used by scientists themselves have sometimes proven insufficient. Below I will discuss these issues, including how research has traditionally been designed and how results have been analyzed.

Questionable Research Practices

Questionable research practices (QRPs) do not reach the level of fraud and were seldom discussed prior to the replicability crisis. They represent a grey area in acceptable scientific practice, and are likely a significant cause of replicability issues in science (Funder et al, 2013, p. 5; John, Loewenstein & Prelec, 2012). QRPs can substantially bias results published in the scientific literature. Simmons, Nelson and Simonsohn (2011) identified several QRPs that are actively used by researchers. They include: strategically deciding when to stop collecting data (i.e., continuing to collect data until the effect has reached statistical significance), using multiple types of measurement but only reporting the successful ones, and failing to disclose experimental conditions in the final report. Simmons et al (2011) performed thousands of simulations to determine the effect of QRPs on the prevalence of false positives. They found that use of all QRPs would cause false positives in experiments 60% of the time.

Beyond simply demonstrating that QRPs can be used to both gain an advantage and introduce false positives into the scientific literature, additional research suggests QRPs are indeed widely used. John and colleagues (2012) performed an anonymous survey of 2,155 research psychologists, asking about their use of QRPs. Over 60% of respondents admitted to selectively reporting a study's measurements and over 45% said they selectively reported studies

that worked. A further 38% reported deciding to exclude data after looking at the impact it would have on the final results (John et al, 2012). Such use of QRPs is probably explained by the fact that prior to the replicability crisis, QRPs were seldom openly discussed, and likely advantageous to a scientist's career. Indeed, a simulation by Bakker and colleagues (2012) finds that the optimal "strategy" in successfully publishing involves employing several small, underpowered studies rather than one large one, and reporting the first one that works. These sobering findings suggest that not only *can* QRPs bias science, they *do*, and at a level higher than most predicted.

The implications for *Daubert*'s testability factor are immense. *Daubert*'s innovation rests on the notion that bad science may be weeded out by asking judges to go beyond general acceptance and evaluate the science itself. But considering that QRPs go nearly undetectable in the published science, there is little to no chance of picking them up under testability. Moreover, considering scientists themselves only recently addressed the issue, it is unlikely that judges would fare any better. This once again illustrates the notion that deficiencies in general accepted practices permeate the rest of the *Daubert* factors, rendering their utility dubious at times. But it is important to note that this conclusion is not fatalistic – the tools exist to manage to QRPs.

Scientists, and psychological scientists in particular, have been laudably proactive in attempting to curb the use of QRPs. As mentioned above, QRPs are now being extensively researched and warned against (Pashler & Wagenmakers, 2012). Prominent researchers and methodologists have provided recommendations to curb QRPs, notably pre-specification of research methodology (e.g., Cumming, 2013, p. 4; Schooler, 2011). For example, millions of dollars have been pledged to the *Open Science Framework* (Open Science Collaboration, 2012), which provides, among other initiatives, an online database for researchers to register

experiments before they are performed, thus reducing the chance that important measurements go unreported. Such pre-registration has long been mandated in clinical research, but remained optional in other scientific fields.

New publication standards have given teeth to the above norms. Beginning in 2014, scientists seeking to publish in *Psychological Science* are *required* to complete a four-item checklist developed based on the norms established by Etienne LeBel and colleagues' PsychDisclosure platform (http://psychdisclosure.org). This checklist involves confirming that the author reported all of the measurements and experimental conditions. It also requires the researchers report all of the excluded data and the reasons for any such exclusion. While this measure still places trust in the experimenters, its binding nature represents a significant step forward in science's commitment to curbing QRPs.

In addition to the mandatory disclosure provisions above, *Psychological Science* has also adopted – although non-compulsorily – the suggestions of the *Open Science Framework*. In particular, the *Open Science Framework* has proposed awarding various "badges" for studies opting to make their data and materials publicly available, as well as for pre-registering the study. As mentioned above, pre-registration is a powerful means of reducing QRPs as it would provide a record of a study's design and methods prior to it being performed and reported, allowing scientists (and potentially judicial actors) to confirm that all conditions or measures were reported. There are, of course, types of research where pre-registration is not practical, a scenario that *Psychological Science*'s systems makes allowances for (Eich, 2013, p. 3) Pre-registration also addresses the file-drawer effect (reviewed above) by providing a record of studies that, for one reason or another were not published, but may provide some practical or

theoretical importance. The badge system, although not mandatory, should motivate researchers to be more open in their practices.

Lack of Replication

> "Confirmation comes from repetition. Any attempt to avoid this statement leads to failure and more probably to destruction." (Tukey, 1969, p. 84)

Few within science would dispute the above quote. Indeed, replication is commonly referred to as the *sine non qua* of scientific research (e.g., see Funder et al, 2013). Unfortunately, despite widespread acknowledgement of its importance, replication is uncommon in general practice. For instance, a recent study of a random sample of published research containing the term "replication" found that only 1.07% of psychological studies are replicated (Makel, Plucker & Hegarty, 2012, p.537). While this finding paints a dim picture of the state of replication in the psychological sciences, the state of affairs in other disciplines is no brighter. For instance, John Ioannidis notes a low level of replication in other scientific disciplines and finds that medical studies often show initially stronger results that regularly prove irreproducible (Ioannidis, 2005a).

One defense of the present regime is that while *direct* replications (i.e., a more-or-less exact duplication of a previous study) are uncommon, *conceptual* replications are relatively frequent. For example, Lykken (1968) suggested that conceptual replications (he termed them "constructive") have a great deal of value in elucidating the construct behind the data. Indeed, conceptual replication involves testing the underlying theory of a study with different methods. For example, a conceptual replication of verbal overshadowing might test whether verbalization impairs memory for other experiences, such as taste or smell. The Makel and colleagues (2013)

study cited above found that of the replications they did find, about 82% were conceptual with the remainder being direct or a mix of the two. While conceptual replication may seem to bolster the believability of a finding, Pashler and Harris (2012, p. 531, 534) suggest it may have what they term "pathological results."

In analyzing direct versus conceptual replications, Pashler and Harris (2012) note that published failed direct replications go a great ways towards increasing understanding of a phenomenon. And even if they are not published, informal channels may help spread the word. Successful conceptual replications, on the other hand, are very likely candidates for publication with unsuccessful ones easily written off as simply pushing the theory too far. It then becomes tempting to publish small conceptual replications and not report the failed ones, thus giving the community the impression a phenomenon is robust. Pashler and Harris (2012, p. 533) therefore conclude that "…entire communities of researchers in any field where direct replication attempts are nonexistent and conceptual replication attempts are common can easily be led to believe beyond question in phenomenon that simply do not exist."

The lack of direct replication is likely a result of lack of incentive. As reviewed above, novelty has long been rewarded in the publication process, and replications are by definition, not novel (Giner-Sorolla, 2012). The most common defense for novelty is that if it is wrong, the science will self-correct (Aronson, 1977, p. 192). Indeed, in a paper comparing legal and scientific values, Gastwirth (1992, p. 56) writes, "Unlike the *Frye* doctrine, science encourages novel experiments and theories because unexpected findings will be reexamined by independent researchers." The empirical evidence reviewed above suggests Gastwirth may have been mistaken as replication has traditionally not been the norm.

As quoted above, confirmation comes from replication. And, above all else, legal decisions depend on accurate input (i.e., evidence). Given the lack of replication of scientific research, it is therefore challenging for judicial decision makers to come to an accurate understanding of some scientific findings and methods. Once again, this does appear to be changing with strong recommendations coming down from scientists and professional bodies encouraging replication. For instance, recommendation six of the *SPSP* Task Force reads, "Encourage and improve the availability of publication outlets for replication studies." (Funder et al, 2013, p. 6)

New Statistics

> "Null-hypothesis significance testing (NHST) has long been the mainstay method of analyzing data and drawing inferences in psychology and many other disciplines. This is despite the fact that, for nearly as long, researchers have recognized essential problems with NHST in general, and with the dichotomous ('significant vs. nonsignificant') thinking it engenders in particular" Eich (2013, p. 3)

Despite a great deal of criticism, much of psychological science and other disciplines continue to rely on NHST as its default method of quantitative argument. In short, NHST allows for the calculation of the conditional probability of the observed data, assuming that the null hypothesis is true. Researchers often set the null hypothesis as the hypothesis that the relationship between two variables is zero. So in other words, null hypothesis significance testing returns the likelihood of the observed data, assuming that in reality there is no effect. This probability is known as a "p-value."

Many have questioned the logic and intuitiveness of NHST over the years (e.g., Carver, 1978; Kirk, 2003). For instance, see the following well-cited passage from Jacob Cohen (1994):

> "What we want to know is 'Given these data, what is the probability that [the null hypothesis] is true?' But as most of us know, what the obtained p-value tells us is 'Given that [the null hypothesis] is true, what is the probability of these (or more extreme) data?'"

Problems with the interpretation of NHST has led both pre- (Wilkinson Task Force, 1999) and post-replicability crisis (Funder et al, 2013) professional bodies to recommend a more fulsome set of statistical procedures. Psychological Science has endorsed one particular set of procedures called "New Statistics" (Cumming, 2013), which represents a movement from NHST to actual estimates of values in question as well as other complementary measures.

One complementary measure of particular use is "statistical power." Calculating and reporting statistical power is, in fact the first recommendation of the SPSP Task force (Funder et al, 2013). Statistical power represents a study's likelihood of rejecting the null hypothesis (i.e., finding an effect), when the null hypothesis is indeed false. In other words, a study with a high degree of statistical power that fails to reject the null hypothesis provides information concerning whether there is indeed no effect because it *could* have found that effect.

Commentators have suggested that running sufficiently powered studies should help curb QRPs (Funder et al, 2013, p. 5). Further, as reviewed above, both law and science are interested in not only the presence of an effect, but the absence. Power helps indicate whether or not a negative finding is indicative of anything at all. In addition, lack of power is associated with several other undesirable results. For example, in what is known as the "winner's curse" phenomenon, low-powered studied tend to overestimate the size an experimental effect (Button

et al, 2013). This is because the low powered studies that do happen to uncover a statistically significant result (i.e., the winners) tend to show aberrantly strong effects because that all these studies are capable of finding. The winner's curse, therefore, results in bias in the publically available data indicating an effect is stronger than it is. Statistical power is not explicitly encapsulated by either the testability or error rate factor, but is an important concept in evaluating the meaning of scientific findings.

Requirements surrounding statistical measures are notoriously difficult to implement as statistics is essentially a form of principled argument (Abelson, 1995), and thus open to multiple interpretations. Still, *Psychological Science* now requires, as part of the checklist described above, that researchers explain how they determined the sample size they used. This determination will frequently involve some discussion of statistical power – indeed, a welcomed advance.

Scientific Evidence in a Post-Replicability Crisis World

If there is one theme to be extracted from the above analysis, it is that the replicability crisis revealed a weakness in the *Daubert* framework. The risk with a general acceptance standard was always that nose-counting can, at times, lead to inaccurate results because the consensus is wrong (Faigman & Monahan, 2009, p. 5). In moving from general acceptance to an analysis of the validity of the science itself, the *Daubert* court did not consider that sometimes those scientific results themselves were the result of inadequate generally accepted practices. The replicability crisis rendered this point painfully salient, and it is one that has important implications for how courts should treat scientific evidence going forward.

Frye or *Daubert*?

Recall that several State courts have refrained from "Daubertizing", instead preferring *Frye's* general acceptance test. This stance is open to criticism on many accounts (e.g., see Faigman & Monahan, 2009; Beecher-Monas, 1998). From a practical standpoint, however, existing scholarship suggests the difference may not matter much at all. For instance, limited scientific fluency among judicial actors likely restricts the effect that *Daubert*'s more thorough analysis can have (McAuliff & Groscup, 2009).

Lessons learned from the replicability crisis suggest that the *Daubert* standard is better equipped than *Frye* to ensure that good science reaches decision makers. As mentioned, the replicability crisis was made possible by deeply entrenched generally accepted practices that did not track with best practices. *Frye*, in probing general acceptance, does not easily get at best practices when such divergences exist. Moreover, identifying and correcting for replicability-related issues will require a more searching review of the science. *Daubert* at least provides a framework that can address the issues described in this Essay. For instance, replication itself easily falls under the *Daubert* head of testing, and concepts from the "New Statistics" fit well under the error rates (Faigman, Monahan, Slobogin, in press). Courts should ensure, however, that *Daubert* is employed in a manner that takes into account best practices for testing, error rates and peer review. Some preliminary suggestions for how this should be done follow a brief review of the impact of replicability issues on various ways scientific evidence may impact legal proceedings.

The Replicability Crisis and Empirical Framework and Diagnostic Expert Testimony

Within expert testimony, a distinction can be drawn between "empirical framework evidence" and "diagnostic evidence" (Monahan & Walker, 1991; Faigman, Monahan, & Slobogin, in press). Empirical framework evidence is that which applies *generally* to the instant

case, providing a context for judicial decision maker to understand an issue at trial. Diagnostic evidence, on the other hand, concerns an application of science to the specific facts of the case. In other words, empirical framework evidence can demonstrate that an effect occurs in the population, whereas diagnostic evidence can demonstrate that it applies to the instant case.

An example is useful in making this distinction between empirical framework and diagnostic evidence more concrete. Much has been written about the hypothetical situation in which fMRI methods are used to detect lies in a legal context (e.g., Schauer, 2009). In this case, an expert providing framework evidence might describe research showing that on average, the brain activity of those lying looks reliably different under an fMRI scanner than those telling the truth. An expert providing diagnostic evidence, however, would be asked to say, based on an fMRI scan if a particular party was lying. This distinction has proven remarkably useful in elucidating different issues that arise in the context of different types of evidence (Faigman, Monahan, & Slobogin, in press). This is no less true for the present Essay.

The replicability crisis is especially problematic for framework evidence. It has been noted elsewhere that with regard to framework evidence, the question of testability is complicated by the variety of methods that can be used to establish a relationship between variables, resulting in a difficult consideration of the weight of the totality of the evidence (Faigman, Monahan, & Slobogin, in press, p. 31). The issues brought up by the replicability crisis complicate this calculus.

The replicability crisis indicates that even noble attempts at surveying a scientific literature with weight placed on findings from well-controlled studies vetted by a prominent journal can lead to miscalculations. For example even a well-meaning examination of the basic

research surrounding fMRI lie detection research, and the academic research linking such research to law (e.g., Schauer, 2009; Stoller & Wolpe, 2007) would have – prior to the replicability crisis – missed a major statistical issue with such research. In particular, recall that Vul (2009) noted that many such studies may underreport their error rates due to statistical issues with multiple non-independent measurements. These issues with basic research can be linked to the peer review issues reported above regarding a lack of attention to methodology and corresponding lack of quantitative experts serving as peer-reviewers.

QRPs also pose a potential problem for framework evidence. It has been noted that "[w]ithout error rate information *Daubert's* command that courts evaluate the reliability of expert evidence cannot be accomplished, at least when empirical framework is involved." (Faigman, Monahan, & Slobogin in press, p. 36). The multiple ways in which error rates can be measured and conceptualized (e.g., *p*-values stemming from NHST, confidence intervals, effect sizes) already presents a challenge for judges. The difficulty with QRPs is that they seriously affect the interpretability of any metric of error rate. With respect to the fMRI lie detection example,[4] it would not be clear from research reported under pre-replicability crisis standards what data or conditions were excluded from the final analysis, thus potentially inflating error rates. As will be suggested below, courts should be wary of QRPs, looking for external indicia of methods that prevent them, and asking experts to report about any conditions or measurements that did not make it to the final analysis.

Faigman, Monahan and Slobogin (in press) note that the concepts of testability and error rate are not always "amenable" (p. 37) to the evaluation of diagnostic evidence. This is because

[4] This example is simply used for convenience. There are no indications I am aware of that such studies contain QRPs beyond concerns with science as a whole.

diagnostic evidence is more concerned with the particular practices and experience of an expert applying a diagnostic technique. Still, framework evidence often provides a context for diagnostic evidence, and they often accompany one another. As such, replicability-related issues will still most likely arise in many cases when diagnostic evidence is given, but more as a threshold matter.

Furthermore, in certain cases, diagnostic standards are girdled by bodies of clinical research that may be probed for replicability-related concerns (Faigman, Monahan, & Sloblogin, 2013, p. 38). In fact, a great deal of debate exists over the extent that clinicians and diagnosticians *should* restrict their practices to that which is science or evidence-based (e.g., see Lilienfeld, 2002; Faigman & Moahan, 2009, p. 22). Still, in many cases, issues stemming from the replicability crisis will be more pertinent with regard to framework evidence.

The Replicability Crisis and Judicial Fact-Finding

The replicability crisis also has implications for what have been termed "legislative facts." These types of facts are those that courts come to when making a decision that goes beyond the dispute at hand (Davis, 1943; Monahan & Walker, 1991, p. 573). A widely cited (Larsen, 2013, p. 102; Monahan & Walker, p. 572) example of legislative fact is Justice Brandeis' factual determination in *Muller v Oregon* (1908) that long working hours were detrimental to female laborers. This type of factual finding was later termed the "Brandeis Brief." Legislative facts may be contrasted to "adjudicative facts," which are applicable only to the case at hand (Monahan & Walker, 1991, p. 576).[5]

[5] Faigman, Monahan, and Slobogin (in press) note that there can be considerable cross-over between legislative fact and adjuctiative fact, instead preferring framework evidence for such material. Here, I focus on legislative facts determined outside of the context of expert testimony brought by parties to the instant case.

Compared to expert evidence proffered by parties to the dispute, the rules and norms concerning admissibility of judicially determined legislative facts are decidedly looser. Larsen (2013, p. 72) notes that legislative facts are explicitly exempted from the *US Federal Rules on Evidence* and indeed advisory notes promote their unrestricted use. Legislative facts can come from a variety of sources, including *amicus* briefs and the judge's own fact-finding endeavors. These facts are regularly based on scientific and social scientific methods. One example is an fMRI study linking video games and aggression cited in Justice Breyer's dissent in *Brown v Entertainment Merchant's Association* (2011) that was subsequently criticized in the scientific community (Larsen, 2013, p. 70; Neuroskeptic, 2011).

Moreover, Larsen recently emphasized the increased precedential value of legislative facts. In particular, she noted that due to the increased ease of keyword searching decisions, it is increasingly easy for lower courts to discover factual findings and cite them as authority. Furthermore, because of the increased length of U.S Supreme Court decisions, there are more of these types of facts to cull from. Larsen argues that this trend presents a serious danger, as these findings of fact may represent a non-veridical portrayal of the science and motivated reasoning on the part of justices (Larsen, 2013 p. 101). In other words, increasing reliance on judicial fact finding can allow poorly vetted scientific finding to have an enduring effect.

Lessons from the replicability crisis are also pertinent for the increasing importance of factual precedent. These findings of fact, when informed by science, are prone to the same replicability-related issues recounted above and may indeed hit harder in the case of factual precedents. For instance, the file-drawer effect is especially problematic. Recall that it is well established that the scientific literature is biased towards positive findings. Expert evidence may indeed provide a check on this phenomenon as the expert may know to refer to meta-analyses

and contact his or her colleagues for unpublished work. It is comparatively unlikely that a justice or clerk will take these corrective measures. Building on this insight, the following section addresses several initiatives previously proposed to improve the legal treatment of scientific evidence, and how they may be best employed in light of the replicability crisis.

Lessons from the Replicability Crisis

As reviewed above, this Essay is far from the first to remark upon the challenges raised as a result science's presence in the courtroom. This work, both empirical (e.g., Gatowski, 2001; McAuliff & Kovera, 2000) and theoretical (e.g., Faigman & Monahan, 2009; Larsen, 2013; McAuliff & Groscup, 2009) is impressive not only for its rigor, but for the tangible recommendations it has produced. In particular, research suggests that court appointed experts, technical advisors and scientific training for judges may be useful in allowing *Daubert* to reach its full effectiveness. After a brief review of these initiatives, I will discuss how they are informed by the lessons learned from the replicability crisis.

Improved scientific training and education for judges may improve the scientific fluency of judges and thus help ensure good science reaches juries (Faigman & Monahan, 2009; Merlino, Dillehay, Dahir, & Maxwell, 2003). Such education could begin in law schools (Merlino et al, 2003) and progress through continuing education programs (Black, Ayala, & Saffran-Brinks, 1994). In Canada, the "Goudge Report" was prepared in the wake of highly publicized cases of flawed forensic pathology affecting several decisions (Goudge, 2008). Recommendation 134 of the report (2008, p. 502) coincides with the prescriptions from the U.S. experience, providing for training programs for judges on a variety of scientific safeguards and the *Daubert* analysis (p. 502). Finally, a wealth of excellent academic guidance is also available as a measure of judicial self-help (e.g., Faigman et al, 2012).

Researchers and commentators have also discussed greater use of court-appointed experts and technical advisors (Faigman & Monahan, 2009; Berger, 1994; Gross, 1991) in assisting with the evaluation of scientific evidence. The use of court-appointed experts is expressly allowed by *Federal Rule of Evidence 706* (2011). In contrast to party-retained experts, the theory with court-appointed experts is that they are more likely to give a non-biased perspective on the science to the jurors. Research is generally mixed over whether these experts tend to prove beneficial or are overly influential on juries (McAuliff & Groscup, 2009). Moreover, there has been substantial judicial resistance to the use of court-appointed experts (Cecil & Willging, 1993). Technical advisors, as compared to court-appointed experts, limit their assistance to judges on matters of scientific evidence. Despite initial misgivings (Faigman & Monahan, 2009, p. 21; Goudge, 2008, p. 506) concerning both court-appointed experts and technical advisors, there appears to be traction for both (Faigman & Monahan, 2009, p. 21).

As an alternative or complement to court-appointed experts, safeguards and procedures surrounding party experts may be imposed. For example, in the context of criminal trials, the Goudge Report (2008, p. 509) recommends pre-trial meetings between opposing experts that may help clarify points of consensus and difference. In such cases the report also recommends a code of conduct to be read and signed by expert witnesses acknowledging a duty to acknowledge the strength of the data supporting an opinion (Goudge, 2008, p. 504). Such procedures have been adopted in England and Wales (*R v Harris*, 2005).

In general, the measures above should be informed by lessons learned from the replicability crisis and stay current with the reforms going on within science. There is a serious risk that history will repeat itself and that several of the new norms and procedures developed in response to the replicability crisis will never be adopted in a binding manner in science. In that

case, it would be up to judicial actors to adopt such safeguards. Fortunately, awareness of the replicability crisis can inspire initiatives that drill down into the aspects of good science that prevent replicability issues. The following recommendations largely follow initiatives that scientists have developed to address the replicability crisis, such as the Society for Personality and Social Psychology's new guidelines (2013), editorial changes at *Nature* (2013) and the *Open Science Framework* (2012).

In response to the replicability crisis, commentators have pointed to the importance of direct replication (Pashler & Harris, 2012; Eich, 2013), or in other words: Has the finding been reproduced by an independent party? This observation has been made before in the legal context (e.g., see Faigman, Saks, Sanders & Cheng, 2006, p. 63) but the replicability crisis indicates that the distinction between direct replication and its cousins can be one that is easy to ignore – even for scientists. Still, prominent methodologists have long considered direct replication to be of hallmark importance in producing valid results (e.g., Cohen, 1994).

While guidance from treatises and professional bodies can certainly be helpful for judges, replication is one area where court-appointed experts can be of special assistance. For instance, direct replications – both those that fail and succeed – are notoriously difficult to publish. Court appointed experts or technical advisors with a background in the field in question would be well-placed to scour the field for evidence confirming or refuting the proffered evidence. Because they possess a background in science, court appointed experts would also be well placed to distinguish between direct and conceptual replications.

With a great deal of unpublished data in existence, as well as data from open-access journals that may not have been well vetted, there is danger of a judge being drowned in

information. In situations such as these, appointing a technical advisor or expert may also be the most sensible choice, as such an individual may be better placed to evaluate a study's methodology and compile several experiments into a meta-analysis to get an idea of the effect size. In fact, meta-analyses have proven to be a useful tool both with respect to assembling information and facilitating understanding of a great deal of data (Schmidt, 1992). While meta-analyses are not without limitations (e.g., Stegenga, 2011), they represent an excellent topic for judicial training programs.

Another factor to consider, when possible, is the degree to which the experimenter disclosed, within reason, all of the variables and measurements he or she used. Openness regarding methodology and procedure is hallmark of good science (Lilienfeld & Landfield, 2008). More specifically, such consideration would help screen out research containing QRPs. An excellent example is the aforementioned *Open Science Framework*, which provides an online database for researchers to register experiments before they are performed, thus reducing the chance that important measurements go unreported. There are, however, several other mechanisms available. The PsychDisclosure platform mentioned above also creates assurances that there were no substantive changes between conception of the experiment and its execution, with the option to provide reasons for any such changes. Furthermore, simple provision of data allows the scientific community to provide independent analysis (Simonsohn, in press).

It would be relatively easy for training initiatives and guides to include information about the availability of preregistration systems (which it should be noted are already a well-accepted component of clinical medical research), and which journals have mandated or recommended them. Clearly, the presence or absence of preregisration is uninformative in the case of most pre-replicability crisis research outside of the medical domain. But in a post-replicability crisis

world, it is an important factor to engage an expert witness about. For example, the badge system, recently endorsed by *Psychological Science* provides an excellent, but optional, external *indicum* of openness of research methods. I don't expect these factors to be relevant in every case, but they are relatively easy to grasp concepts and if contemporary research failed all of these *indicia* of openness, it may raise red flags about the research.

Openness of research methods also impacts *Daubert*'s peer review and publication factor. As reviewed above, research indicates that jurors do seem swayed by the publication status of research (Kovera, McAuliff & Hebert, 1999). Therefore it would be useful for judges to examine not just whether a study has been peer reviewed, but how that peer review was conducted. Once again drawing from the openness considerations, it stands to reason that a journal that has adopted such standards will have engaged in a higher level of vetting. On the other end of the spectrum, there is reason to think that the bad reputation some open access journals have for performing a cursory peer review process is a justified reason to discount their findings (Bohannon, 2013).

Daubert's peer review factor has been considered little more than a logical extension of general acceptance (Beecher-Monas, 1998). But as compared to testability and error rate, it is a factor that is easy to grasp by lay audiences. Ease of understanding cuts both ways, rendering peer review a factor that presents both a danger of overvaluation, but one ripe for training initiatives. Indeed, guides and training programs prepared by academics but aimed at judges could easily convey information about the degrees of quality of peer review, including their demands for sharing of research material, data, disclosure and preregistration. It would also be simple and nearly costless to maintain an online database of various publications and the level of scrutiny they represent with regard to replicability concerns (e.g., those that have adopted post-

replicability crisis measures, to those that failed to identify bait experiments such as in Bohannon's study in *Science* noted above).

On the other hand, it will not always be clear how thorough the peer review itself was. This is especially true with respect to the quantitative aspects of review. For instance recall *Psychological Science* made a statement embracing the statistical movement away from NHST (Eich, 2013). Further, *Nature* (Editorial, 2013) declared it would assign methodological and quantitative experts to review research when appropriate. While these declarations are important, they will be difficult to assess unless and until they are deployed in a transparent manner. At this stage of science's recovery from the replicability crisis, such mechanisms are not yet in place.

In light of the difficult to assess nature of statistics and methodology, in cases where such complicated issues arise, judges would be wise to be liberal in appointing a technical advisor or court-appointed expert. Moreover, in jurisdictions where procedural law allows for it, court-mandated pre-trial meetings between party experts could be useful in honing in on these issues. In fact, research has demonstrated (Brown, Tallon, & Groscup, 2008, described in McAuliff & Groscup, 2009) that when specifically instructed to educate the jurors, the presence of opposing experts can assist jurors' understanding of the expert evidence. Therefore, if experts are focused in on a specific methodological issue and are under a code of conduct to educate, it may be that challenging issues will be dealt with in a manner that provides for both the strength and weaknesses of both parties' positions. In addition when methodological issues are not serious enough to merit a finding of inadmissibility, judges may wish to issue special instructions to jurors regarding the weight to be assigned to evidence.

In summary, law's mission going forward will be to ensure that lessons learned regarding science's best practices go on to influence judicial decision making. *Daubert* presents a sensible framework whereby testing, error rate, and peer review can be drilled down into to more specific replicability-promoting concepts (e.g., direct replication, rigorous peer review, new statistics). How these concepts are worked into judicial practice largely depends on the complexity of the issue and the resources available to judges.

Conclusion

While much of the above discussion sounds pessimistic, the outlook within science is quite hopeful. Even before recent revelations, a great deal of science was performed and reported with care and rigor. And with tangible changes such as the work being done by the *Open Science Framework* and the editorial changes at leading academic outlets, there is reason to believe that bodies of work will only become more reliable. For instance, the *Open Science Framework* recently released its first report (Klein et al, 2013) on the replication attempts of several leading psychological theories. This effort found that many effects were reproducible by independent laboratories, with only a few (and indeed those that had already attracted some apprehension, see Yong, 2012) resting on uncertain ground (e.g., Carter, Ferguson & Hassin, 2011; Caruso et al, 2013).

The large scale self-correction science is undertaking is indeed heartening. But while science involves abstractions – theories that can withstand periods where the evidence self-corrects – law is not allowed the same liberties. Litigation requires immediate findings of fact, and as such the replicability crisis provides an important cautionary tale. It is possible that history will repeat itself, and the norms being developed in response to the replicability crisis never become widely adopted. In that case, to ensure the highest quality of science impacts legal

proceedings, it will be up to the legal profession to ensure that lessons from the past are not forgotten. I am optimistic about the capability of judicial actors in adopting these best practices. *Daubert* may not have been designed with the replicability crisis in mind, but it provides a framework that judges – with proper training and guidance – can follow to account for most replicability-related issues.

References

Abelson, R. P. (1995). *Statistics as a Principled Argument*. New Jersey: Psychology Press.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Aronson, E. (1977). Research in Social Psychology As A Leap of Faith. *Personality and Social Psychology Bulletin, 3,* 190-192.

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7,* 543.

Beecher-Monas, E. (1998). Blinded by Science: How Judges Avoid the Science in Scientific Evidence. *Temple Law Review, 71,* 55-102.

Bem, D.J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology, 100,* 407.

Berger, M. (1994). Novel forensic evidence: The need for court-appointed experts after *Daubert*. *Shepard's Expert and Scientific Evidence*, 1, 487.

Bernstein, D. E., & Jackson, J. D. (2004). *Jurimetrics, 4,* 351.

Black, B., Ayala, F. J., & Saffran-Brinks, C. (1994). Science and the law in the wake of *Daubert*: A new search for scientific knowledge. *Texas Law Review, 72,* 715-802.

Bohannon, J. (2013). Who's Afraid of Peer Review? *Science, 342,* 60.

Bower, B. (2012). The Hot and The Cold of Priming. *Science News, 181*, 26.

Brown, M., Tallon, J., & Groscup, J. (2008, March). The effectiveness of opposing expert testimony as an expert reliability safeguard. In J. Groscup (Chair), *Scientific reliability in the courtroom: Jurors' assessments of scientific evidence and legal safeguards*.

Symposium conducted at the annual meeting of the American Psychology-Law Society, Jacksonville, FL.

Brown v Entertainment Merchants Assocation, 131 S. Ct. 2729, 2768 (2011) (Breyer, J., dissenting).

Button, K. S., Ioannidis, J. P. A., Mokryz, 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.

Carey, B. (2011). Fraud Case Seen as Red Flag for Psychology Research. *The New York Times, A3*.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, *22(8)*, 1011-1018.

Caruso, T. J., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, *142*, 301-307.

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

Cheng, E. K., & Yoon, A. H. (2005). Does Frye or Daubert Matter? A Study of Scientific Admissibility Standards. *Virginia Law Review*, *91*, 471-513.

Cecil, J. S., & Willging, T. E. (1993). Court-Appointed Experts: Defining the Role of Experts Appointed Under Federal Rule of Evidence 706 (pp. 83-88). Washington, DC: Federal Judicial Center.

Chesebro, K. J. (1994). Taking Daubert's "Focus" Seriously: The Methodology/Conclusion Distinction, *Cardozo Law Review*, *15*, 1745.

Chin, J. M. & Schooler, J. W. (2008). Why Do Words Hurt? Content, Process, and Criterion Shift Accounts of Verbal Overshadowing. *European Journal of Cognitive Psychology, 20*, 396.

Christophersen v Allied-Signal Corp., 939 f. 2d 1106 (5th Cir. 1991).

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

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all?: Rethinking some conclusions of the literature on judgment under uncertainty. Cognition, 58, 1-73.

Coursol, A. & Wagner, E. (1986). Effect of Positive Findings on Submission and Acceptance

Rates: A Note on Meta-Analysis Bias. *Professional Psychological Research Practices, 17,* 136.

Cumming, G. (2013). The New Statistics: Why and How. *Psychological Science, 24*, 1.

Davis, K. C. (1942). An Approach to Problems of Evidence in the Administrative Process. *Harvard Law Review, 55*, 364.

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579-595. (1993).

Dickersin, K. (1990a). Publication Bias and Clinical Trials. *Controlled Clinical Trials, 8,* 343.

Dickersin, K. (1990). The Existence of Publication Bias and Risk Factors for its Occurrence. *JAMA, 263*, 1385.

Editorial. (2013). Announcement: Reducing Our Irreproducibility. *Nature, 496,* 398.

Eich, E. (in press). Business Not as Usual. *Psychological Science*.

Faigman, D. L. (1999) *Legal Alchemy: The Use and Misuse of Science in the Law*. New York: W. H Freeman and Company.

Faigman, D. L., Blumenthal, J. A., Cheng, E. K., Mnookin, J. L., Murphy, E. E., & Sanders, J. *Modern Scientific Evidence: The Law and Science of Expert Testimony (Vols. 1-5).* Eagen, MN: Thomson/West.

Faigman, D. L. & Monahan, J. (2009). Standards of Legal Admissibility and Their Implications Psychological science. In J. L. Skeem, K. S. Douglas, & S. O. Lilienfeld (Eds.), *Psychological Science in the Courtroom Consensus and Controversy* (pp. 3-25). New York, NY: Guildford.

Faigman, D. L., Monahan, J., & Slobogin, C. (in press). Group to Individual (G2i) Inference in Scientific Expert Testimony. *University of Chicago Law Review*, *81*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2298909.

Faigman, D. L., Sanders, M. J., Sanders, J., Cheng, E. K. (2007). *Modern Scientific Evidence: Forensics*. Eagen: West Academic Publishing

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891.

Federal Rules of Evidence Rule 702 (2011).

Federal Rules of Evidence Rule 706 (2011).

Fox, R. E. (2000). The dark side of evidence-based treatment. *Practioner Focus*.

Friedman, R. D. (2003). Minimizing the Jury Over-Valuation Concern, *Michigan State Law Review,* 967.

Frye v. United States. 293 F. 1013, 1014. (D.C. Cir. 1923).

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. G. (2013). Improving the Dependability of Research in Personality and Social Psychology: Recommendations for Research and Educational Practice, *Personality and Social Psychology Review, 17,* 1.

Gatowski, S. I. (2001). Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-Daubert World. *Law and Human Behaviour, 25,* 433.

Galak, J. (2012). Correction the Past: Failure to Replicate Psi. *Journal of Personality and Social Psychology, 103,* 933.

Gastwirth, J.T. (1992). Statistical Reasoning in the Legal Setting. *The American Statistician, 46,* 55-56.

General Electric Co. v Joiner. 522 U.S. 136. (1997).

Giner-Sorolla, R. (2012). Science or Art? Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science, 7,* 562-567.

Goudge, S. T. *Inquiry into Pediatric Forensic Pathology in Ontario*. Retrieved from http://www.attorneygeneral.jus.gov.on.ca/inquiries/goudge/index.html.

Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin, 82,* 1.

Groscup, J. L., Penrod, S. D., Studebaker, C. A., Huss, M. T., & O'Neil, K. M. (2002). The effects of *Daubert* on the admissibility of expert testimony in state and federal criminal cases. *Psychology, Public Policy, and Law, 8,* 339.

Gross, S. (1991). Expert Evidence. *Wisconsin Law Review,* 1113.

Ioannidis, J. P. A. (2005a). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA, 294,* 218.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Med. e124, 2*.

Jasonoff, S. (1996). *Science at the bar: Law, science, and technology in America*. Cambridge, MA: Harvard University Press.

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the Prevalence of Questionable

Research Practices With Incentives for Truth Telling. *Psychological Science, 23,* 524.

Johnson, G. (2014). Truths Only One Can See. *The New York Times, D1*.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83-105). Malden, MA: Blackwell.

Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., Galliani, E. M., et al., (2013). Investigating variation in replicability: The "Many Labs" Replication Project. Retrieved from Open Science Framework, osf.io/wx7ck.

Kovera, M. B., & McAuliff (2000). The effects of peer review and evidence quality on judge evaluations of psychological science: Are judges effective gatekeepers? *Journal of Applied Psychology*, *85*, 574-586.

Larsen, A. O. (2014). Factual Precedents. *University of Pennsylvania Law Review, 162,* 59.

Lehrer, J. (2010). *The New Yorker*. Retrieved from http://www.newyorker.com/reporting/ 2010/12/13/101213fa_fact_lehrer.

Likwornik, H. (2011). Bayes Rules – A Bayesian-Intuit Approach to Legal Evidence. (Unpublished doctoral dissertation). University of Toronto, Toronto. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/32076/1/Likwornik_Helena_M_201111_PhD_thesis.pdf.

Lilienfeld, S. O. (2002). The scientific review of mental health practice: Our raison d'etre. *Scientific Review of Mental Health Practice*, 5-10.

Lilienfeld, S. O., & Landfield, K. (2008) Science and Pseudoscience in Law Enforcement A User-Friendly Primer. *Criminal Justice and Behavior*, *35*, 1215.

Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2004). *Science and pseudoscience in clinical psychology*. New York: Guildford.

Mahoney, M. J. (1977). Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System. *Cognitive Therapy and Research, 1*, 161.

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012). Replications in Psychological Research: How Often Do They Really Occur? *Perspectives on Psychological Science, 7*, 537.

McAuliff, B. D., & Groscup, J. L. (2009. *Daubert* and Psychological Science in Court. In J. L. Skeem, K. S. Douglas, & S. O. Lilienfeld (Eds.), *Psychological Science in the Courtroom Consensus and Controversy* (pp. 26-52). New York, NY: Guildford.

McAuliff, B. D. Kovera, M. B., & Nunez, G. (2009). Can jurors recognize missing control groups, confounds, and experimenter bias in psychological science? *Law and Human Behavior*, *33*, 247.

McNutt, M. (2014). Reproducibility. *Science*, *343*, 229.

Meissner, C. A. & Brigham, J. C. (2001). A Meta-Analysis of the Verbal Overshadowing Effect in Face Identification. *Applied Cognitive Psychology, 15,* 603.

Merlino, M. L., Dillehy, R., Dahir, V., & Maxwell, D. (2003). Science education or judges: What where, and by whom? *Judicature*, *86*, 210.

Monahan, J. & Walker, L. (1991). Judicial Use of Social Science Research. *Law and Human Behavior*, *15*, 571.

Muller v Oregon, 208 US 412 (1908).

Neuroskeptic (2011). Violent Brains in the Supreme Court. Retrieved from http://neuroskeptic.blogspot.ca/2011/07/violent-brains-in-supreme-court.html.

Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science, 7,* 657.

Pashler, H. & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science, 7,* 531-534.

Pashler, H. & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?. *Perspectives on Psychological Science, 7,* 528.

R v Harris and others, [2005] EWCA Crim 1980.

R v JLJ, [2000] 2 SCR 600.

R v Mohan, [1994] 2 SCR 9.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE, 7(3): e33423*.

Robinson, E. (2011). Not Feeling the Future: A Failed Replication of Retroactive Facilitation of Memory Recall. *Journal for the Society of Psychical Research, 75*, 142.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance For Null Results. *Psychological Bulletin, 86,* 638.

Sanders, J. Diamond, S. S., & Vidmar, N. Legal Perceptions of Science and Expert Knowledge. *Psychology, Public Policy, and Law*, *8*, 139.

Schauer, F. (2009) Can Bad Science be Good Evidence? Neuroscience, Lie Detection, and Beyond. *Cornell Law Review*, *95*, 1191.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Pscyhologist, 47*, 1173.

Schooler, J. W. (2011), Unpublished results hide the decline effect. *Nature, 470*, 437.

Schooler, J. W. & Engstler-Schooler, T. (2000). Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid. *Cognitive Psychology*, *22*, 36.

Schwartz, A. (1997). A 'Dogma of Empiricism' Revisited: *Daubert v. Merrell Dow Phamaceuticals, Inc.* and the Need to Resurrect the Philosophical Insight of *Frye v. The United States*. *Harvard Journal of Law & Technology*, *10*, 149.

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*, 1359.

Simonsohn, U. (in press). Just Post it: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence?. Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences, 42(4), 497-507.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn From Tests of Significance – or Vice Versa. *Journal of the American Statistical Association, 54*, 30.

Sterling v Velsicol Chem. Corp. 855 F. 2d 1188 (6th Cir. 1988).

Stoller, S. E., & Wolpe, P. R. (2007). Emerging Neurotechnologies for Lie Detection and the Fifth Amendment. *American Journal fo Law and Medicine*, *33*, 359.

Stroebe, W., Postmes, T. & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science, 7*, 670.

Tukey, J. W. (1969). Analyzing Data: Sanctification or Detective Work. *American Psychologist, 24*, 83-84.

United States v Hall. 974 F. Supp. at 1202.

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

Wagenmakers, E. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem. *Journal of Personality and Social Psychology, 100*, 426-431.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data is Related to the Strength of the Evidence and the Quality of Reporting Statistical Results. *PLoS ONE e26828, 6*.

Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54*, 594.

Yong, E. (2012). A Failed Replication Attempt Draws a Scathing Personal Attach From a Psychology Professor. *Discover Magazine*. Retrieved from http://blogs.discovermagazine. Com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen.

Yong, E. (2012). Replication Studies: Bad Copy. *Nature, 485*, 298-300.