A Manifesto for Open Science in Giftedness Research

Matthew T. McBee[1]

Matthew C. Makel[2]

Scott J. Peters[3]

Michael S. Matthews[4]

[1]Department of Psychology, East Tennessee State University, Johnson City, TN

[2]Duke University Talent Identification Program, Durham, NC

[3]Department of Educational Foundations, University of Wisconsin-Whitewater, Whitewater, WI

[4]Department of Special Education and Child Development, University of North Carolina-Charlotte, Charlotte, NC

Abstract

The ruinous consequences of currently accepted practices in study design and data analysis have revealed themselves in the low reproducibility of findings in fields such as psychology, medicine, biology, and economics. Because giftedness research relies on the same underlying statistical and sociological paradigms, it is likely that our field also suffers from poor reproducibility and unreliable literature. This paper describes open science practices that will increase the rigor and trustworthiness of gifted education's scientific processes and their associated findings: open data; open materials; and preregistration of hypotheses, design, sample size determination, and statistical analysis plans. Readers are directed to internet resources for facilitating open science. A preprint of this paper is available at [redacted].

Keywords: pre-registration, reproducibility crisis, p-hacking, QRPs, HARKing, open science, open data, open materials, registered report

Word Count: 8,362

A Manifesto for Open Science in Giftedness Research

Introduction

The social sciences are currently experiencing a crisis of confidence (Pashler & Wagenmakers, 2012). This crisis affects the very foundations of scientific research: reproducibility, generalizability, and accuracy. If a field cannot reproduce results, generalize them to other situations or contexts, or have confidence in the accuracy of the results, then what contribution has been made? How has a scientific approach moved the field forward? The origins and causes of these problems are longstanding and complex, but are important to understand if the general process of scientific practice is to be improved. Using how all these issues relate specifically to gifted education research as a backdrop, we devote the first section of the manuscript to introducing readers to many of the problems of the modern research process, then the second section introduces readers to some of the principles of open science that we believe will help gifted education researchers produce research with more reproducible, generalizable, and accurate results.

We begin with an example from social psychology, where numerous supposedly well-established phenomena have not been reproduced in replication studies. For example, in the famous "pen study", Strack, Martin, and Stepper (1988) asked participants to hold a pen either between their teeth (activating the same muscles used while smiling) or between their lips (activating muscles used while frowning) and rate the humor of a set of Far Side comics. In the original study, comics were rated significantly funnier when participants held the pen between their teeth. This finding was used to support the facial feedback theory of emotion that facial movements *cause* rather than respond to emotional states. At the time of this writing (April, 2017), the Strack et al. paper had been cited 1,590 times (as counted by Google Scholar) and is

discussed in numerous introductory psychology textbooks. As far as the field was concerned, this finding was factual. However, in August 2016 the result of a large-scale, preregistered, multi lab collaborative replication of the original study was published in *Perspectives on Psychological Science* as part of their Registered Replication Report (RRR) series (Wagenmakers, Beek, Dijkoff, & Gronau, 2016). In this project, seventeen independent research teams replicated Strack et al.'s original experiment. The project leaders then performed a meta-analysis over all seventeen of the results to produce a single aggregated finding. The result showed no evidence of a facial feedback effect, which was estimated as 0.03 ± 0.13 for a 95% meta-analytic confidence interval of [-0.11, 0.16]. In fact, according to a Bayesian analysis, all 17 labs produced evidence supporting the null hypothesis of zero effect, with the vast majority providing reasonably strong evidence for the null (e.g., $BF_{01} > 3$). Figure 1 contains a forest plot from this study, illustrating the results across the participating labs as well as the meta-analytic result. It is clear from this plot that there is no consistent effect of facial feedback on humor ratings. This is just one of several recent examples where seemingly established findings have not been reproduced (e.g., Open Science Collaboration, 2016).

Insert Figure 1 here

Problems with replication extend beyond social psychology. In recent years, the low rate of replication of research has become a major concern (c.f., Makel, Plucker, & Hegarty, 2012; Makel & Plucker, 2014). In gifted education, Makel and Plucker found only one in 200 articles claimed to have attempted any kind of replication. The lack of attempted replications means false positive findings in our literature are rarely discovered, challenged, or corrected--in most cases

the status quo is considered to be settled once initial information is obtained. For example, in the U.S., 13 states required a nomination to identify gifted students (McClain & Pfeiffer, 2012). However, in recent research Grissom and Redding (2016) found that teacher nominations for gifted identification can exacerbate the gaps between identification rates of white, black, and Hispanic students. This happens even though one of the original reasons for using nominations was to increase the diversity of students being identified! When research claims or educational practices are not evaluated, the existence or consistency of the hypothesized effects will be unknown. Researchers in psychology have responded to such issues by substantially increasing their replication efforts. Unfortunately, no such response has yet occurred in education; one goal of this brief is to help introduce possible paths that can facilitate such a response.

When effects are nonexistent in reality, how is it that researchers find evidence for those effects in in the first place? All statistical hypothesis tests have the potential for a false positive (Type-I) error, in which the researcher concludes that there is statistically significant evidence of a nonexistent effect. In fact, that primary virtue of the so-called *null hypothesis significance testing* (NHST) paradigm is Type-I error control. The risk of making such an error is controlled by two things--the choice of the alpha criterion (by tradition, set at the 5% level), and more importantly, the process that was used to obtain the statistical evidence. This last point is sometimes referred to as *researcher's intentions* (Goodman, 1999; Kruschke, 2012; McBee & Field, 2017; see also Rouder et al., 2016). Although the first is often reported and well understood, the second point is also crucial yet much harder to evaluate as far as its effect on Type-I error rates.

Sociological Context for Research. The research process is embedded in a sociological context that strongly incentivizes particular behaviors. Behavioral incentives aren't necessarily

problematic, but deep problems arise when they encourage counterproductive or damaging behaviors. For example, most research is performed by university faculty and graduate students. These individuals are evaluated on their research productivity, as measured by the number of publications produced and the prestige of the journals in which they are published. Individuals must demonstrate research productivity in order to be competitive for academic jobs and for grant funding. Assistant professors need publications to get tenure and be promoted, and associate professors need publications and often grant funding to be promoted to full professor. These incentives play out at the university level, as much of the current model of higher education financing at research-intensive universities rests on overhead costs funded through grant revenue, connecting research productivity closely to the financial health of the institution. Under this incentive model, quantity of production is highly prized. More is better. Fewer (but better quality) publications typically have a lower value in such a system.

At the same time, academic journals have a limited amount of space in which to put content--their "page budget"--and this traditionally has been dictated by printing and binding limitations[1]. The prestige of journals is related to their circulation and readership in a field, and this often is measured by the number of times their articles are cited within other articles in the same field, as measured primarily by the Impact Factor calculated by publisher Thomson Reuters.

Working within these constraints, journals are incentivized to publish research that will garner the most citations and thereby increase the journal's prestige and impact factor. As a result, journals seek to publish work that is exciting, new, that tells a compelling, 'clean' story, and above all, that finds statistically significant evidence supporting the central claims. Despite

[1] Obviously this is an anachronism given that online dissemination has largely replaced print as the method by

the central value of null findings in a falsificationist model of science (Dienes, 2016), non-significant findings often are viewed as uninteresting. They therefore are less likely to be published, or cited if they do get published; because peer reviewers and journal editors are less willing to publish non-significant findings, so null results often end up in the researcher's file drawer (Sterling, 1959) where they are inaccessible to the research community. As a result, the published literature provides a biased sample of the universe of research that was actually undertaken[2].

As Schimmack (2012) pointed out, this publication bias is, in itself, sufficient in the long-run to destroy the error control properties of *p*-values within the NHST framework. In other words, when only statistically significant findings are published, the rate of false positives among the *published* literature is much higher than the nominal 5% that researchers assume to be the case. For example, if nonexistent phenomenon X is studied 20 times, one would expect one of these tests to yield a false positive Type-I error. But if this study is the only one published on phenomenon X, the evidence base in the literature will seem to support the existence of X. Fanelli (2010) found that 91% of papers in psychology and psychiatry reported evidence supporting the focal hypothesis, the highest of any field he selected. This suggests either that psychologists have such penetrating theoretical insight into human behavior that their hypotheses are almost never wrong, or that publication in psychology is strongly biased toward statistically significant findings. We leave consideration of the relatively likelihood of these explanations as an exercise to the reader. Moreover, the low acceptance rate of psychology and education journals means that competition for the limited slots, dictated by the limitations of outdated print

[2] When the literature has a bias against null findings, research synthesis methods such as meta-analysis produce incorrect estimates of effects (see discussion in Engber, 2016). Popular bias-correction techniques such as funnel plots and PET-PEESE probably do not correct these problems (Simonsohn, 2017).

dissemination, is intense. Researchers who wish to have (or keep) careers are *strongly* incentivized to find streamlined, internally consistent, and statistically significant stories to tell. Their livelihoods depend on it.

Unfortunately, many of the incentives researchers face do not encourage good science. Publication bias is bad enough, but it is not the sole problem. Most researchers are aware of the multiple comparisons problem. When researchers test a set of hypotheses, the risk of making at least one false positive decision increases dramatically with the number of tests[3]. Because statistical significance is practically required for publication in many areas of the social sciences, and because researchers are evaluated on the basis of publications, it follows that researchers are strongly incentivized to produce statistical significance. Luckily for them, but unluckily for meaningful scientific progress, statistical significance is not hard to produce so long as a sufficient number of (unadjusted) statistical tests are examined. Given how easy computers have made it to conduct hundreds if not thousands of statistical tests, this is a realistic problem. When the published paper fails to describe the multiple comparisons that led to the result, but rather presents the isolated test that 'worked', readers are given a *highly* misleading impression regarding the strength of the evidence supporting the study's claims. This is known as the multiple comparisons problem.

Strategies for converting nonsignificant findings into statistically significant findings (likely to be Type-I errors, of course) are collectively known as *p-hacking* and these appear to be among the most widespread of the various questionable research practices (QRPs; Simmons, Nelson, & Simonsohn, 2011). A common set of such strategies were discussed in Simmons et al.'s now-classic piece entitled, "False-Positive Psychology: Undisclosed Flexibility in Data

[3] The risk is given by (1-(1-α)^k) where k is the number of tests. This quantity is called the familywise error rate.

Collection and Analysis Allows Presenting Anything As Significant." The four strategies that they discussed were (1) measuring additional outcome variables, (2) adding 10 additional participants per condition and rerunning the hypothesis tests, (3) controlling for a covariate, and (4) dropping (or including) experimental conditions. When these four strategies were used in combinations, the result was a 60.7% false positive rate. The researchers then applied these strategies to real data and found statistically significant evidence that listening to the Beatles song "When I'm Sixty-Four" *literally* resulted in participants becoming older. John, Lowenstein, and Prelec (2012) surveyed a group of 2,000 psychologists about their use of these (and other) p-hacking strategies such as fraudulently rounding p-values below the .05 threshold and falsifying data. Their results indicated that these practices are quite commonly used. For example, 66.5% of their respondents indicated that they had dropped non-significant outcomes from their papers, and 58% reported engaging in 'optional stopping' (Lakëns, 2014), in which the decision to collect more data is conditioned on a non-significant interim statistical test. QRPs are far from uncommon, and there's no reason to suspect that gifted education research has been insulated from such practices. To quote Simmons, Nelson, and Simonsohn's (in press) retrospective on their false-positive psychology paper: "Everyone knew it was wrong, but they thought it was wrong the way it's wrong to jaywalk. We decided to write 'False-Positive Psychology' when simulations revealed it was wrong the way it's wrong to rob a bank."

Another questionable research practice that bears mention is *hypothesizing after the results are known* (HARKing; Kerr, 1998). HARKing often occurs in response to a 'failed' confirmatory study -- one in which the primary hypotheses were unsupported by statistical evidence. Having obtained an answer that they do not like, and that will not facilitate career

progress due to file drawer bias, researchers search for a question to which the answer is 'yes.' (Makel, 2014, p. 4).

One common response to a failed confirmatory study is to convert it after the fact to an exploratory study instead. For example, researchers who fail to observe a hypothesized main effect of treatment may then examine unplanned post-hoc subgroup analyses. If an overall effect wasn't found, perhaps a treatment effect will be observed in males only or in African American students, or in Title I schools but not others. Given the reality of confirmation bias and the ease with which the human mind can concoct ex post facto explanations for almost anything, it is quite easy in most cases to produce a plausible theoretical explanation for these post-hoc findings. Based on our combined experience reviewing journal and conference submissions in the field of gifted education, we believe this is one of the most common QRPs in the field. Perhaps a gifted curriculum showed no main effect on student learning, but when authors conducted several rounds of post-hoc subgroup analyses, they found a statistically significant effect for low-income, African American students. The curriculum is then presented as if it is a research-supported intervention for low-income African Americans when, in reality, this finding is likely to be completely spurious.

HARKing becomes especially problematic when researchers take the extra step of rewriting the whole paper, including the literature review and hypotheses, around these incidental findings. In this case, an exploratory study is presented as though it had been confirmatory all along. These two modes of research produce markedly different levels of evidence, and in our view it is fraudulent to misreport findings in this way. Unfortunately, current norms do not prohibit such practices, and in practice, current incentives nearly demand them. Few researchers have the luxury of 'wasting' a study by not publishing it when it could be

salvaged using techniques that are not prohibited, and in some cases are even requested by journal editors and peer reviewers. It is hard to insist that researchers behave properly when, as Simmons, Nelson, and Simonsohn (2012) wrote, "there is no shared understanding of what 'properly' is" (p. 4-5). In the absence of strong norms about the boundaries of acceptable practices, grey areas, and fraud, HARKing and other QRPs will continue to flourish. During the peer review process, it is not uncommon for reviewers to even suggest that authors *change* their hypotheses in a revised manuscript.

Many researchers believe that HARKing (and other QRPs, such as optional stopping) are fully responsible for the findings reported in Bem's (2011) infamous precognition paper. This paper was published in the *Journal of Personality and Social Psychology*, a flagship social psychology journal, and in it were presented nine studies seeming to show that the future causes the past (Wagenmakers, Wetzels, Boorsboom, & van der Maas, 2011). For example, in Bem's Experiment 1, participants could predict with greater-than-chance accuracy (p=.01) on which side of the screen an erotic image would appear, but not any of the other four types of images that were examined. It strains credulity past the breaking point to believe that a specific hypothesis predicting precognition effects for erotic stimuli but no other type existed in advance of seeing the data, and the pattern of sample sizes across Bem's nine experiments is suggestive of optional stopping (Yarkoni, 2011). Indeed, the publication of this paper is now considered to be one of the seminal events in psychology's crisis of confidence, as it provided an extraordinarily clear example of the fallibility of current statistical and research practices (Gelman, 2016). If status quo methods could produce such implausible findings, then it stands to reason that they could have produced erroneous findings many times before and since -- and in

ways that were insidiously undetectable to peer reviewers or journal editors, the supposed quality control agents of academe.

Although examples from outside gifted education may seem remote to some readers, we have noted that many papers in gifted education articulate research questions rather than hypotheses. The term "research questions" implies that there were no predictions or *a priori* hypotheses. Without predictions, a research project cannot be confirmatory; it can only be exploratory. The confirmatory-exploratory distinction is critical. It is only in confirmatory research that test statistics (e.g., *t* tests, *F* tests , and *p* values) can be used, as these are inappropriate for exploratory research (de Groot, 1969; Lakëns & Evers, 2014). As McBee & Field (2017) wrote, "reporting *p*-values (and confidence intervals) is a privilege, not a right" (p. 64). This is not to say that exploratory research is not valuable, only that generating questions is not the same as providing answers. The difference is theory building as opposed to theory testing.

Given the reporting standards that do not mandate that authors describe the true process that led to their claims, perverse incentives operating at the level of individual researchers and academic journals, widespread adoption of *p*-hacking strategies, and lack of a robust culture of replication (meaning that false positive claims are never detected), one might predict that the overall truth value of published claims in the social sciences and allied fields is very low. And indeed, that is precisely what has been observed. The Reproducibility Project: Psychology (RP:P) was a collaborative attempt by 270 individual researchers to replicate 100 studies published in a variety of subdisciplines of psychology. 97% of the original studies reported statistically significant effects with a mean effect size of .403. On the other hand, only 36% of the replications reported statistically significant results, with a mean effect size of .197. The

authors deemed that 39% of the original effects were successfully replicated (Open Science Collaboration, 2015). Three of the four large-scale Registered Replication Reports published in *Perspectives* failed to obtain any evidence in favor of the original claim. In the field of cancer biology, Begley and Ellis (2012) attempted to replicate a set of 59 'landmark' studies in the field. These papers had been published in journals with impact factors of at least five (larger than nearly any education journal) and had, on average, several hundred citations each. Only 11% of these studies were reproducible. Neither the number of citations nor the impact factor of the publishing journal had any relationship with reproducibility, which again points to the strong incentives at work. Multiple studies now suggest that there is no relationship, or possibly even a negative one, between the prestige (via impact factor) of a journal and the reproducibility of its contents. For example, Figure 2 was taken from Ziemann, Eren, and El-Osta (2016) and displays the rate of erroneous gene names published in genetics papers by journal. Note that *Nature* had the highest error rate (IF=38.14) while *Molecular Biology and Evolution* (IF=13.65) had the lowest.

Insert Figure 2 here

Open Science

In response to many of the problems described above, a growing movement in the social sciences revolves around a range of practices generally referred to as open science. Listing all the benefits of open science practices is beyond the scope of the current manuscript (for a more complete review, see Makel & Plucker, 2017), but open science not only helps improve the quality of the research produced, but is also believed to benefit the individual researchers who

use such techniques (e.g., Markowetz, 2015; McKiernan et al., 2016; Wagenmakers & Dutilh, 2016). Below, we introduce four open science practices (preregistration, open data, open materials, and preprints/figure sharing), what each contributes to the research endeavor, and how individual researchers can begin incorporating them into their practices. Table 1 summarizes some open science practices, the problems they help to address, and resources to assist in their implementation.

Insert Table 1 about here

Preregistration

Preregistration requires researchers to articulate their hypotheses and analyses plans *prior to* data collection. Several websites, such as the Open Science Framework (OSF; osf.io) and aspredicted.org, provide free and easy ways for researchers to preregister their predictions and analysis plans (including a step by step how-to example that can be found: https://osf.io/sgrk6/). It is a commonly held misperception that preregistration creates additional work for authors. In fact it only shifts work; it doesn't add work. Researchers have to write these sections at some point anyway (often for IRB approval); preregistration simply shifts the timeframe in which the writing is done. Use of a third party website allows researchers to demonstrate time-stamped evidence of what their starting hypothesis was and when it was preregistered. This simple action can provide readers and reviewers with confidence that any HARKing, p-hacking, or other QRPs were not the cause of a finding. Importantly, these plans can be kept completely private until after the research has been completed so that others cannot see what researchers have preregistered. This privacy prevents others from scooping or even plagiarizing others' good

ideas. Thus, pregistration offers a very straightforward way to increase the validity and overall trustworthiness of a body of work.

Misconceptions about preregistration. There are numerous widely stated concerns about preregistration that are misleading at best. Preregistration does not hinder exploratory research or frown upon its use. Rather, preregistration helps differentiate what is exploratory from what is confirmatory research. Acknowledging that the two types of research are different is not a burden; it is a necessary part of the scientific endeavor. Moreover, after preregistering hypotheses and analysis plans, researchers can still add subsequent analyses that were developed after data collection. Preregistration simply helps make it clear which is which. In this way any reader easily can determine which tests were confirmatory and which were exploratory.

Benefits of preregistration. Preregistration is easy, free, and increases the credibility of the work by providing evidence that researcher degrees of freedom were limited. The only "downside" is that preregistration does not allow *p*-hacking and/or choosing a path through the garden of forking paths that leads to the statistical conclusion that the researcher wants to reach. Having evidence showing that neither of these was done increases the credibility and rigor of a manuscript as well as the larger body of knowledge. However, it should be noted that if what is preregistered is vague and or incomplete, it will provide little support to bolster the manuscript. In their checklist on avoiding p-hacking, Wicherts and colleagues (2016) note that preregistration provides strength only to the extent that specificity, precision, and exhaustiveness are included.

The contributions that preregistration adds to researchers and the research process are numerous. Wagenmakers and Dutilh (2016) recently outline "seven selfish reasons for pregistration" for psychology researchers, but the benefits they suggest can be generalized to researchers in other domains as well. These include giving clear credit for actual predictions (and

providing evidence that they aren't HARKed), adding protection against post-publication accusations, and helping avoid the de facto criteria of requiring results to be statistically significant in order to be published in an academic journal. Most published findings in the social sciences represent statistically significant effects (Fanelli, 2010; 2012), but the fundamental goal of scientific research is not statistical significance or even novelty; it's the pursuit of ever-greater approximations of the underlying truth. Removing the burden of requiring statistical significance from published research will help remove harmful incentives for individual researchers to engage in QRPs by rewarding asking important questions and relying on strong methods instead of supporting sensational findings based on weak methods. Moreover, given that research papers are fundamentally persuasive essays whose goal is to convince readers to update their beliefs about or understanding of the world (McBee & Field, 2017), it is in the author's direct personal interest to take actions that increase the credibility, convincingness, or perceived truth value of their work. Credibility is a prerequisite for impact, and we believe that preregistration is the single most effective way that authors can increase the credibility of their claims.

The benefit that preregistration provides to the larger field are also numerous. Peer review is fallible; it misses mistakes and can be easily manipulated, even without ill intent. Avoiding *p*-hacking increases trust in results by decreasing the amount of trust required of the individual researcher. For example, when clinical trials started requiring preregistration of hypotheses, "success" (as in non-null findings) dropped from 57% to 8% of trials (Kaplan & Irvin, 2015). Many current researchers might find this disturbing since, under the current incentives, it means they will publish fewer studies, but it also means the published studies are more likely to be a true representation of the phenomenon under investigation.

Open Data and Open Materials

Open data is the public sharing of (de-identified) data collected and analyzed as part of a research study. Similarly, open materials refers to the sharing of research materials (e.g., IRB proposal text, consent forms, survey items, analysis code) with others. Open data and open materials can be shared on an institutional site or via a third-party site (e.g., https://osf.io/).

Benefits of open data and open materials. Openness and transparency are fundamental scientific values. The Royal Society, established in 1660, is perhaps the world's oldest scientific society. Its motto of *Nullius In Verba* translates roughly as "Take No One's Word." Science should not rely on trust, but rather should facilitate the open verification of findings and claims. And verification requires that individuals be able to access the materials needed to check every link of the inferential chain for errors or fraud, and also that the procedure for reaching a set of conclusions is fully documented. *Open data* means that researchers publicly post their (deidentified) data online to a permanent repository upon article submission. *Open materials* means that any auxiliary materials, including statistical analysis code, lab notes, data collection schedules, and instruments are publically shared. Ideally, any interested party should be able to run the researcher's analysis code on their publicly-posted data and reproduce every value, statistical test, table, and figure in a manuscript. As Daniel Lakëns (2017) wrote, "When you want to evaluate scientific claims, you need access to the raw data, the code, and the materials." Readers are simply unable to determine the validity of scientific claims without access to this information.

Open data and open materials have numerous benefits (*Psychological Science* Editorial, 2017). For individual researchers, these can serve as an archive and backup of all data and materials, providing a record of all previously used materials that is not tied to a specific computer or employer's infrastructure, thus preventing "lost" files. Open data and open materials

also allow other researchers to use previously used data sets and materials for their own research questions (including for research syntheses and meta-analyses), thus putting the data to more use without putting an extra burden on researchers to respond to requests or find files they may not have accesses in years. Moreover, rather than making such materials available only upon request, making them open implies that they are available automatically; no request must be made and there is no bureaucracy limiting access to the materials. Such default sharing benefits the field by saving future researchers time from re-inventing the wheel, as well saving them from having to request materials that other researchers may not have used for several years or even decades.

Collecting data and creating materials takes time and resources, so their creators should get credit for this work. For example, while re-using someone else's analysis code can be a huge time saver, the person who created that code also deserves credit. Citing prior publications is the most commonly used form of giving credit in current practice. But sharing materials actually increases the number of ways in which creators can be given credit.

Open materials are helpful resources, but these should not be assumed to be without fault. Mistakes get made and do not always get caught by individual authors or even in the peer review process. Thus, all researchers should thoroughly evaluate all previously used materials, regardless of whether they have been used previously.

Beyond the OSF, there are numerous cost free resources for sharing data and materials. The only cost to users is the time spent in learning how to use and integrate them into their research practices. For example, at a basic level, shared online accounts through programs such as Google Drive and Google Documents can be used to share materials among co-authors as well as to make this process public. Figshare allows sharing data, figures, and even entire

manuscripts. Another similar option, Github, facilitates open source and collaborate development of code that can be shared from the onset or upon completion.

Preprints

Preprints are published drafts of research manuscripts that are posted online prior to having gone through the traditional journal review process. Some fields, such as physics through the website arXiv.org, have a long tradition of posting preprints as soon as manuscript are complete and then going through the traditional journal review process. arXiv started hosting preprints in 1991 and now hosts over 1.2 million of these documents. More recently, several other fields have created similar sites, such as PsyArXiv.org (psychology), SoArXiv.org (sociology), and more generally, osf.io/preprints, which is an open preprint repository that allows for search of preprints across domains. Preprints posted to these services are now indexed by Google Scholar and other scholarly search tools, enabling interested readers to discovery and access these works.

Benefits of preprints. Preprints remove the delay between study completion and publication. The current academic publication system juxtaposes the steps of research evaluation and research publication, while separating these steps via preprints helps accomplish several relevant goals. First, preprints help reduce the file drawer effect (Rosenthal, 1979), as well as other malign incentives that may encourage QRP, by allowing authors to share all of their results regardless of their significance, clarity, or novelty. Second, preprints accelerate the dissemination of findings by moving dissemination ahead of external evaluation. By allowing the authors (instead of others) to determine when their work is ready to be shared, researchers are given greater autonomy over both their content (in terms of *what* and *how* this content is presented) and the timing of its publication and dissemination. Third, preprints allow authors to share their work

in a format that does not require all readers to either have a paid subscription to a journal or to pay to access the specific content. Moreover, preprints also allow individual researchers to establish copyright ownership of the figures they create. Finally, preprints help simplify the roles of journal reviewer and editors by removing the gatekeeping role and focusing efforts on the goal of evaluation of the research (Nosek & Bar Anon, 2012). To find out if a journal you plan to submit to accepts pre-prints, you can visit their webpage or http://www.sherpa.ac.uk/romeo/search.php (but check to make sure this is up to date for your journal of interest).

Actions Journals Can Take

In order to address the concerns described above, there are several steps journals can take. Below we describe these steps and offer suggestions for their implementation, should journal editors and publishers find these concerns compelling enough to do so.

Open Science Badges

Badges are small icons that appear on the title page of published articles. These communicate to readers that the article in question has been produced in accordance with particular open science practices Badges offer a near zero-cost (to both the researcher and the journal) mechanism for recognizing and rewarding open science practices. Put simply, badges are a form of recognition for research that has met standards for open science, and there is empirical research that they serve to increase the use of open science practices and to increase the availability and accuracy of data (Kidwell et al., 2016). The most common badges (see Figure 3) relate to open data, open methods, and preregistration.

Insert Figure 3 about here

How such badges serve to improve the quality of research is simple. Either when an author first submits an article for publication, or when he or she submits a final version (after acceptance), he or she can indicate which open science practices have been utilized in the study. For each badge, the author answers a series of questions. For the open data badge, these questions are as simple as indicating the permanent online link where researchers can find the original data, on which the analyses were based, for replication purposes. Such a permanent link is also provided such that other researchers can locate the methods used in sufficient detail in order to conduct a true replication. This might include software code or simply a more-detailed version of a data analysis section. The journal itself may, but need not, house these data or methods descriptions. As noted earlier, several sites already exist that can do this at no charge. For the preregistration badge, the author provides evidence of the preregistration time stamp along with assurances that the final analyses match those presented in the preregistration. The author's answers and assurances are then reviewed by the journal and, if they are satisfactory, the article is published with the relevant badges pictured on the first page of the article. This conveys to the reader a higher level of review and trustworthiness in the methods and resulting findings. Over time, removing the potential for *p*-hacking, HARKing, or other QRPs will increase the overall quality of the knowledge base within a given field.

Badges indicate that the arguments being made in the paper are not as contingent upon trust as is usually required. Instead, key aspects of the evidence supporting the paper's arguments are independently verifiable. For this reason, badges signal quality and trustworthiness. We believe that it is likely that badged articles will be cited more frequently as well. At the very

least, the findings presented in badged articles are more likely to influence the field and persuade skeptics (McBee & Field, 2017). And isn't that what it's all about?

Registered Reports

An additional option that journals can take is to allow for the submission of registered reports for consideration rather than only accepting completed manuscripts. Registered reports involve the author submitting only the introduction, literature review, hypotheses, and proposed methods for peer review, prior to any data collection or analysis. This initial stage of a registered report is quite similar to a grant dissertation proposal. The journal reviewers then assess the quality of the study design and the information it will provide, without any consideration of the desirableness of the findings. A well-designed study should not be reviewed less favorably simply because it showed a particular gifted education intervention was ineffective. Instead, it should be reviewed and accepted or rejected based on its motivation, hypotheses, and methodological quality, and this is what registered reports seek to do. If successful in its proposal, the project receives an in-principle acceptance for publication. Another round of review occurs after data are analyzed, but this is more perfunctory than the traditional review process. The registered report model completely removes any incentive for researchers to engage in *p*-hacking. For more detail on registered reports, see Chambers, Feredoes, Muthukumaraswamy, and Etchells (2014); Nosek and Lakëns (2014); or https://cos.io/rr/#RR.

In addition to dramatically improving the quality and trustworthiness of the scientific literature, the registered reports format offers tremendous benefits to researchers as well. The current peer review process evaluates completed projects. It is very often the case that reviewers identify major shortcomings in a study's design or instrumentation that admit alternative explanations of the findings (in other words, threats to internal validity). At this point in the

research process, it is too late to alter the flawed design because the data have already been collected. In some cases, this leads to the study being rejected for publication altogether, in which case the authors have wasted considerable time, resources, and effort. Research subjects may have been exposed to potentially harmful interventions for nothing. In other cases, authors, having been rejected by a "flagship" journal, will resubmit the paper to a different journal that is perceived to be less selective. The paper is once more sent out for peer review, encumbering another 3-4 experts with uncompensated work. If the paper is eventually deemed to be publishable, it is certain that the authors will need to add extensive discussion of the study's flaws to the Limitations section of the manuscript. The paper ends of having less impact and receives fewer citations that the authors had hoped because, in the final analysis, the evidence for the study's central claims is weak.

The registered reports format is a far more humane process for researchers and reviewers than the current system. Peer review of the study's methodology occurs at a point in the process when changes can actually be made, not after it is too late. Reviewers and editors are no longer put in the difficult position of telling their colleagues that a study is doomed and that effort was wasted. Nor are reviewers placed in the unenviable position of second-guessing a study's purported claims (e.g., "*there's no way that effect size can be so large!*") due to unprovable suspicions about the use of QRPs. The registered reports format, if adopted by our field as an option, would result in a much more efficient use of labor, research subjects, and peer reviewers. These are valuable bonuses that go above and beyond the primary benefit of better, more reliable science. We daresay that individual researchers would experience a considerable improvement in the quality of their professional lives and a decrease in stress. At the time of this writing (april 2017), 49 journals (for an updated list, see: https://cos.io/rr/#journals) are now accepting

registered report submissions, and the list is growing on a weekly basis. We hope that all of the gifted education journals will soon appear on this list.

A Call to Action

In our ideal world, all journals would incentivize open science practices (leaving room for reasonable exceptions). That said, we know some of these changes would greatly challenge the existing incentive system and, as such, would require time and other resources. Here are several simple steps a journal can take to move toward more open practices:

1. Move immediately to implement the badge system. This would involve minimal work for a journal, its reviewers, or its editors. Journals could add checkboxes to their submission system whereby authors could indicate if they have preregistered their study, if they will make all data available via a permanant link, or if they will make all methods (in sufficient detail to allow for replication) available via a permanent link. Articles not wishing to commit to open science methods could still submit and be reviewed, but those who do wish to move in this direction are able to do so and have the badges printed on their articles. This will provide readers with a greater degree of confidence in the results of the paper as well as inform them if data and methods are immediately available for replication or additional research purposes.
2. Encourage reviewers to look for open science practices when reading manuscript submissions. Many reviewers already look to see if study methods are presented in sufficient detail to allow for replication. Another step a journal could take would be to encourage its reviewers to look for statements about preregistration, more detail on the specific hypotheses being tested and if the data will be made publically available. Even if the badge system is not incorporated into a journal, reviewers can still emphasize these

practices in a ways that can begin to change the culture around open science practices. The Peer Reviewer Openness initiative (https://opennessinitiative.org) is one means of facilitating this goal. Signatories of this initiative agree to make openness a requisite component of the peer review process.

3. Clarify exploratory versus confirmatory practices. Editors and reviewers can crack down on inferential methods being used for exploratory purposes. In there is no a priori hypothesis and no clear population to which the results are to be inferred, then inferential statistics are inappropriate.
4. Over time, move toward incentivizing open science practices. For example, perhaps preregistered studies receive priority review in the journal's cue, or that studies seeking badged for any open science practices are guaranteed a certain turn around time in review. In gifted education journals, those accepted articles that have adhered to open science practices could receive special attention at annual conferences, could compete with each other for a new award, or could be made available in an open-access format in order to broaden the availability of the work. There are endless ways to encourage authors to engage in open science practices.
5. Accept and encourage replication and null findings. Editors should make clear that they support and welcome replication research submissions for publication. Further, editors should be on the lookout for reviewers who are hostile toward a submission simply because it found a nonsignificant effect or because it was replicating a prior study. As Makel and Plucker have argued (2014), facts are more important than novelty.
6. Allow for authors to submit research for review before the results are known via the registered reports format. Very little of the review process would need to change.

Reviewers would not be able to be influenced by whether or not the findings were positive for a particular agenda, or if they were statistically significant or not. Moreover, authors would get feedback on their method and analysis plans at a time when they can actually ethically alter these plans and avoid mistakes they may have overlooked.

7. Encourage that authors derive, when possible, specific numeric predictions from theories ("point predictions") to test against data, rather than testing a null hypothesis of zero effect -- which typically bears little relation to the theory being tested, and may not even be credible.

Before moving on from this topic, we want to again point out how low cost and low effort most of these open sciences practices really are. Committing to a research hypothesis before conducting analyses doesn't require extra time on the part of reviewers or editors beyond accepting a preregistration certificate. Checking to make sure data sets have indeed been uploaded to a permanent repository would similarly take only a few moments. We believe the benefits to the field enormously outweigh the costs.

Discussion

A field that seeks to be at the cutting edge of education can also reap great benefits from being at the cutting edge of research practices. This may be doubly true for fields such as gifted education that often struggle to garner policy support. Every additional aspect of methodological rigor that can be added to research armors the subsequent results from detractors. Open science practices such as preregistration, open data, open materials, and preprints can help improve the rigor of research, and increase access to important materials, which can help disseminate quality work.

No research study is perfect, but that does not mean that methodological rigor cannot improve the value individual studies provide. None of the practices discussed here are a panacea. Nor do they combine to avoid all problems that can arise in the research process. However, they each help reduce or remove common ailments that weaken research results. Such efforts will increase the credibility of the field's work (Munafo et al., 2017) and help us not just accelerate our students, but also our own understanding of our students.

References

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

De Groot, A. D. (1969). Methodology: foundations of inference and research in the behavioral sciences. The Hague: Mouton.

Editorial. (2017). Sharing data and materials in Psychological Science. *Psychological Science*. Retrieved from: http://journals.sagepub.com/doi/pdf/10.1177/0956797617704015

Engber, D. (2016). Everything is crumbling. *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.html

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904.

Gelman, A. (2016). What has happened down here is the winds have changed [blog post]. Retrieved from http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Goodman S. N. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine, 130*, 995-1004. doi: 10.7326/0003-4819-130-12-199906150-00008

Ioannidis, J. P. A. (2005c). Why most published research findings are false. *PLoS Medicine, 2*, 696–701. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Lakëns, D. (2017). Five reasons blog posts are of higher scientific quality than journal articles [blog post]. Retrieved from http://daniellakens.blogspot.com/2017/04/five-reasons-blog-posts-are-of-higher.html

Lakëns, D. & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278-292. DOI: 10.1177/1745691614528520

Kruschke, J. K. (2013). Bayesian estimation supercedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573-603. doi: 10.1037/a0029146

Makel, M. C. (2014). The empirical march: Making science better at self-correction. *Psychology of Aesthetics, Creativity, and the Arts, 8*, 2-7. DOI: 10.1037/a0035803

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43*(6), 304-316. doi: https://doi.org/10.3102/0013189X14545513

Makel, M. C., Plucker, J. A., Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7,* 537-542. doi: https://doi.org/10.1177/1745691612460688

Makel, M. C. & Plucker, J. A. (2017). (Eds). *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*. Washington DC: American Psychological Association.

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology, 16*, 274-279. DOI 10.1186/s13059-015-0850-7

McBee, M. & Field, S. (2017). Confirmatory study design, data analysis, and results that matter. In M. C. Makel and J. A. Plucker (Eds). *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research.* Washington DC: American Psychological Association.

McKiernan, E., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A. Lin, J., … & Yarkoni, T. (2016). Point of view: How open science helps researchers succeed. eLife, 5:e16800. **DOI:** http://dx.doi.org/10.7554/eLife.16800

Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., … & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*, DOI: 10.1038/s41562-016-0021

Pashler, H., & Wagenmakers, E. J. (2012). Introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530. doi:10.1177/1745691612465253

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*, 520-547.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*(4), 551-566. Doi: 10.1037/a0029487

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*, 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (in press). False-positive citations. *Perspectives on Psychological Science*.

Simonsohn, U. (2017). [58] The funnel plot is invalid because of this crazy assumption: r(n.d)=0. [blog post]. Retrieved from http://datacolada.org/58

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*, 30–34. doi:10.2307/2282137

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*(5): 768–777. doi:10.1037/0022-3514.54.5.768

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Mass, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426–432.

Wagenmakers, E. J. & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer*, retrieved from: http://www.psychologicalscience.org/observer/seven-selfish-reasons-for-preregistration#.WDxeGNUrLRZ

Wicherts, J. M.Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1832. doi: 10.3389/fpsyg.2016.01832

Yarkoni, T. (2011). The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong [blog post]. Retrieved from: http://www.talyarkoni.org/blog/2011/01/10/the-psychology-of-parapsychology-or-why-good-researchers-publishing-good-articles-in-good-journals-can-still-get-it-totally-wrong/

Figure 1

Forest plot of facial feedback replication findings from Wagenmakers, Beek, Dijkoff, & Gronau (2016)

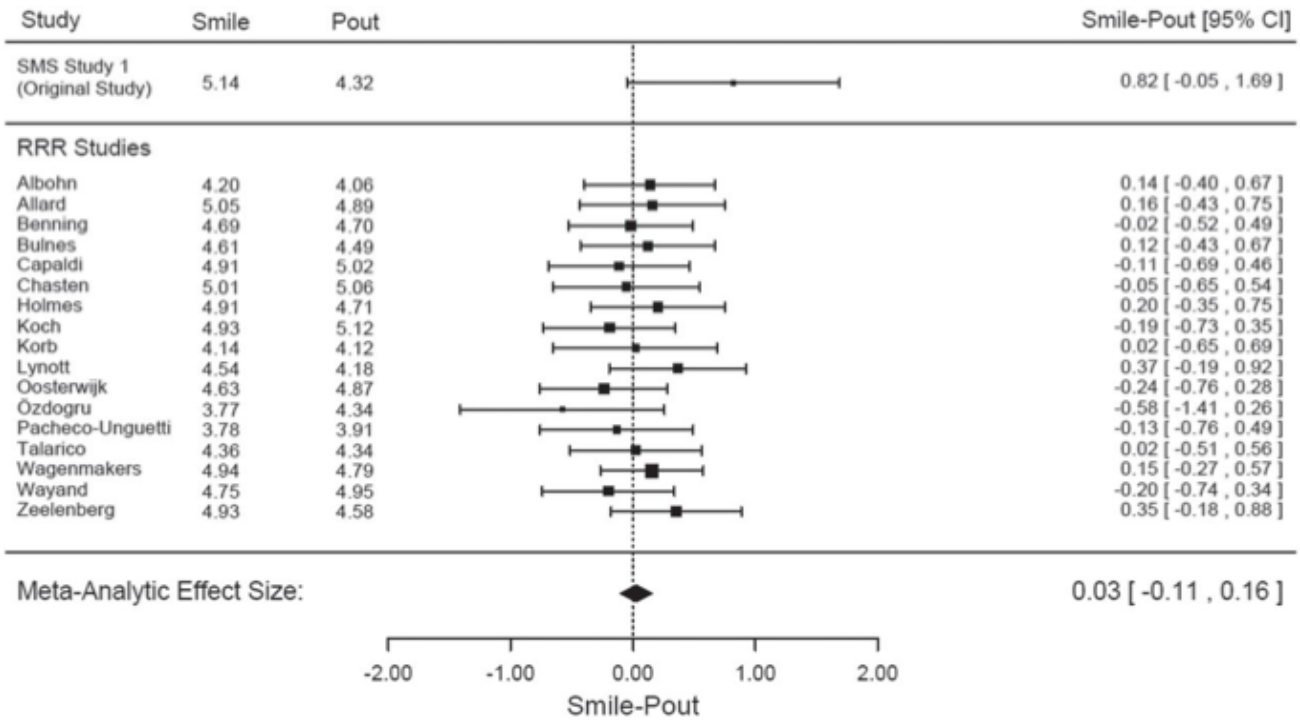

Figure 2

Gene name error rate by journal from Ziemann, Eren, & El-Osta (2016)

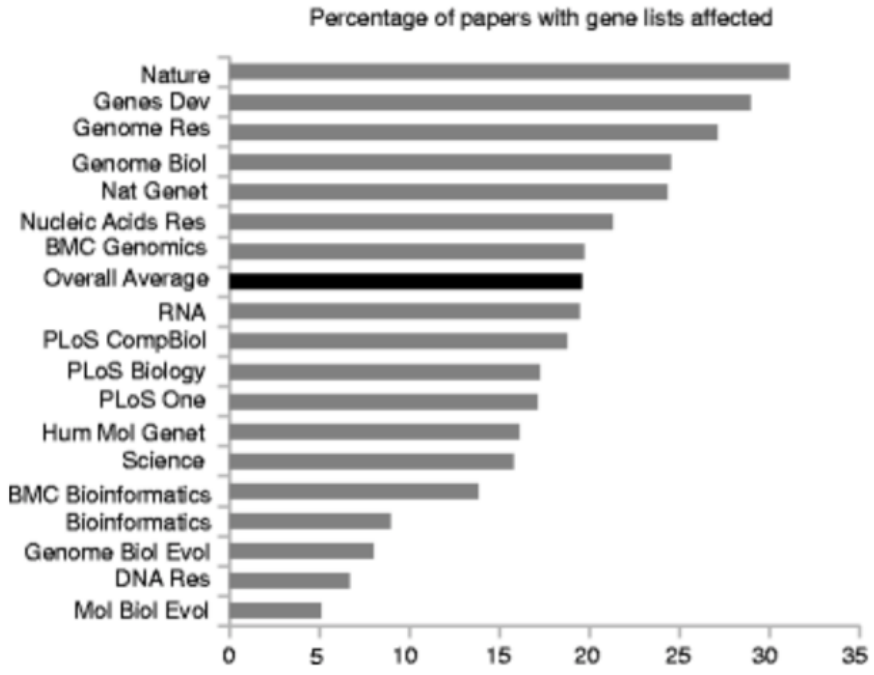

Figure 3

Badges acknowledging open science practices

Table 1

Resources for implementing Open Science practices

Problem(s)	Solution	Benefits	Resources
Lack of rigor, reproducibility, transparency, openness in research	Open science practices	Greater ability to predict, generalize, reproduce research findings	Center for Open Science: https://cos.io/
HARKing, p-hacking, selective reporting of findings	Preregistration	Allows researchers to establish credibility of rigor and predictions; helps avoid post-publication critique	Open Science Framework: osf.io and its how to: https://osf.io/sgrk6/ As Predicted: aspredicted.org
Lack of long-term archive for data and materials, difficulty accessing previously used data and materials	Open Data & Open Materials	Archive of data and materials that can be accessed by anyone without an intermediary; facilitates replication and verification	Open Science Framework; http://osf.io
Lag between completion and publication, conflation of publication and evaluation	Preprints	Immediate dissemination; Reduction of file drawer effect	PsyArXiv.org
Lack of incentives/recognition for open science practices	Badges	Badges currently exist for: preregistration, open data, and open materials. Badges provide acknowledgement and incentives for using open science practices	List of journals using badges and how to begin using badges: https://osf.io/tvyxz/wiki/5.%20Adoptions%20and%20Endorsements/
Publication bias favoring sensational and/or statistically significant findings over important questions	Registered Reports	Helps researchers avoid pitfalls before data are collected; Refocuses research on importance of method and the questions asked;	Journals accepting registered reports: https://cos.io/rr/#journals