Facts Are More Important Than Novelty: Replication in the Education Sciences

Matthew C. Makel and Jonathan A. Plucker

EDUCATIONAL RESEARCHER published online 13 August 2014
DOI: 10.3102/0013189X14545513

FEATURE ARTICLE

Facts Are More Important Than Novelty: Replication in the Education Sciences

Matthew C. Makel[1] and Jonathan A. Plucker[2]

Despite increased attention to methodological rigor in education research, the field has focused heavily on experimental design and not on the merit of replicating important results. The present study analyzed the complete publication history of the current top 100 education journals ranked by 5-year impact factor and found that only 0.13% of education articles were replications. Contrary to previous findings in medicine, but similar to psychology, the majority of education replications successfully replicated the original studies. However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles. The results emphasize the importance of third-party, direct replications in helping education research improve its ability to shape education policy and practice.

Keywords: assessment; content analysis; educational policy; evaluation; replication; research methodology

Carl Sagan (1997) once wrote,

> At the heart of science is an essential balance between two seemingly contradictory attitudes—an openness to new ideas, no matter how bizarre or counterintuitive they may be, and the most ruthless skeptical scrutiny of all ideas, old and new. This is how deep truths are winnowed from deep nonsense. (p. 304)

The desire to differentiate "truth from nonsense" has been a constant struggle within science, and the education sciences are no exception. Over the past decade, these efforts have been especially pronounced at the federal level, with the creation of the Institute for Education Sciences (IES) within the U.S. Department of Education. More to the point, IES funded the creation of the What Works Clearinghouse (WWC) in 2002 to serve as a "central and trusted source of scientific evidence for what works in education" (WWC, 2013). Similar, alternative resources have been created, including the Doing What Works website, which draws largely from the WWC, and the IES-supported Center for Data-Driven Reform in Education's Best Evidence Encyclopedia.[1]

The efforts of these and related initiatives rely heavily on randomized controlled trials (RCTs) and meta-analyses. For example, although the structure and activities of the WWC have evolved over time, a RCT design was originally required for inclusion in the WWC. RCTs were colloquially deemed the "gold standard" in education research by IES (e.g., Whitehurst, 2003). In a similar vein, a National Reading Panel (1999) report stated,

> To make a determination that any instructional practice could be or should be adopted widely to improve reading achievement requires that the belief, assumption, or claim supporting the practice can be causally linked to a particular outcome. The highest standard of evidence for such a claim is the experimental study, in which it is shown that treatment can make such changes and effect such outcomes. (p. 1-7)

Of course, RCTs—and even meta-analyses of RCTs—have their limitations, as do all approaches to research. Problems that can tarnish the gold standard include, but are not limited to, bias toward publishing positive findings (e.g., Bakker, van Dijk, & Wicherts, 2012; Bozarth & Roberts, 1972; Fanelli, 2010, 2012; Pigott, Valentine, Polanin, Williams, & Canada, 2013; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995), low reliability among peer reviewers (e.g., Cicchetti, 1991; Cole, 1992; D. Peters & Ceci, 1982), hypothesizing after the results are known (e.g., N. Kerr, 1998; Makel, 2014; Maxwell, 2004), misuse of statistical tests and results (e.g., Bakker & Wicherts, 2011: Kruskal &

[1]Duke University, Durham, NC
[2]University of Connecticut, Storrs, CT

Educational Researcher, Vol. XX No. X, pp. 1–13
DOI: 10.3102/0013189X14545513

Majors, 1989), the file drawer problem of nonpublished studies that arrive at negative or mixed findings (e.g., Rosenthal, 1979; Rotton, Foos, Vanmeek, & Levitt, 1995; Spellman, 2012), the decline effect from large initial findings (e.g., Ioannidis, 2005a; Rhine, 1934/1997; Schooler, 2011), overreliance on null hypothesis testing (e.g., Bakan, 1966; Cohen, 1994; Johnson, 2013; LeBel & Peters, 2011; Lykken, 1968; Rozeboom, 1960; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), experimenter/researcher degrees of freedom (e.g., Rosenthal, 1966, 1967; Simmons, Nelson, & Simonsohn, 2011), and data peeking (e.g., John, Loewenstein, & Prelec, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Given the potential weaknesses of experimental designs, there is ample evidence of the need to go beyond the gold standard. Indeed, Campbell and Stanley (1963), in their seminal chapter on the use of experimental designs in education research, cautioned against considering true experiments to be a panacea, suggesting instead that they be viewed as a path to accumulating understanding.

One additional avenue to accumulate understanding is replication. Broadly, replication is the purposeful repetition of previous research to corroborate or disconfirm the previous results. Replications also comprise the research used to compose meta-analyses. However, it is important to note that meta-analyses are not the same as replications (Makel & Plucker, 2014). Replication is necessary for meta-analysis, but meta-analyses can be based on studies with quite varied purposes. For example, a meta-analysis on the effects of academic acceleration could rely on studies investigating grade skipping and early entrance into kindergarten even though the individual studies would not be considered replications of each other. Thus, studies may come from the same meta-analytic pool but not serve the same purpose. Meta-analyses synthesize previous research, whereas replications seek to verify whether previous research findings are accurate. Additionally, meta-analyses do not easily account for biases against reporting all outcomes (see Pigott et al., 2013), whereas such biases can be uncovered via replication.

Other fields within the social sciences, most notably, psychology, in recent years have found that replication not only helps pave the path to understanding but also serves as a way to reveal fraud. In this paper, we review classic and contemporary thinking about replication, note current debates about replication in other fields, and provide data on the current state of replication in the education sciences.

Conceptions of Replication

Despite being one of the basic building blocks of science, there is no universally agreed upon list of necessary and sufficient features of a replication (for an overview, see Schmidt, 2009). In a classic paper on how to interpret statistical significance, Lykken (1968) stated that researchers "are interested in the construct . . . not in the datum" (p. 156). To organize these interests, he proposed three types of replications: literal, operational, and constructive. Literal replications "involve exact duplication of the first investigator's sampling procedure, experimental conditions, measuring techniques, and methods of analysis" (Lykken, 1968, p. 155). This form of replication basically calls for the original investigator to collect more data from additional participants in a consistent manner; literal replications are often considered to be impossible (Hansen, 2011; Lindsay & Ehrenberg, 1993; Madden, Easley, & Dunn, 1995) or to suffer from the very same experimenter bias that replications attempt to address. Operational replications are when the researcher "strives to duplicate exactly just the sampling and experimental procedures" (Lykken, 1968, p. 155). Independent researchers can (hopefully) follow the same "experimental recipe" from the original Methods section. Finally, in constructive replications,

> one deliberately avoids imitation of the first author's methods. . . . One would provide a competent investigator with *nothing more than* a clear statement of the empirical "fact" which the first author would claim to have established . . . and then let the replicator formulate his own methods of sampling, measurement, and data analysis. (Lykken, 1968, pp. 155-156; italics in the original)

Operational replications test the veracity of the original data, whereas constructive replications test the targeted construct.

A recent review by Schmidt (2009) connects replication theory with replication in practice. Schmidt lists five functions replications serve: to control for sampling error, to control for artifacts, to control for fraud, to generalize to different/larger populations, or to assess the general hypothesis of a previous study. Rather than deliberately avoiding the original methods, Schmidt suggests systematically changing individual facets of the original study to better understand its nature. The review also reframed replication into direct and conceptual replications. Similar to Lykken's (1968) first two replication types, direct replications repeat the experimental procedure, whereas conceptual replications use different methods to test the underlying hypothesis. We use Schmidt's conceptualization in this paper.

The relative importance of direct and conceptual replications has been debated. Some scholars argue that conceptual replication should be emphasized (e.g., Levy, 1969; Loevinger, 1968; Ostrom, 1971; Smith, as cited in Yong, 2012), while others support direct replications (e.g., LeBel & Peters, 2011; Ritchie, Wiseman, & French, 2012). The importance of each depends on the goal of the investigation (Jones, Derby, & Schmidlin, 2010; La Sorte, 1972; Rosenthal, 1969), with direct replication typically seeking to verify or corroborate the original findings using the same methods as the original researchers; conceptual replications test more general models and theories. However, it is important to note that only direct replications can disconfirm or corroborate previous claims. This is because a failed conceptual replication does not automatically identify a flaw in the original study but instead has the potential to identify the generalizability (or lack thereof) of the original finding. Direct replication, on the other hand, can help identify potential biases in the original study or confirm that the original finding was not an anomaly. Because of this, some argue direct replication should always precede conceptual replication attempts (e.g., Pashler & Harris, 2012).

Why Not Replication?

Replication research can help identify, diagnose, and minimize many of the methodological biases listed above, with Collins

(1985) going so far as to call replication the Supreme Court of science. Despite the benefits that replication brings to the research table, conducting replications is largely viewed in the social science research community as lacking prestige, originality, or excitement (Lindsay & Ehrenberg, 1993; Neuliep & Crandall, 1993b), a bias that is not always shared in the natural sciences (Madden et al., 1995, but cf. Bissell, 2013). Several recent publications have begun to discuss the hurdles and disincentives to conducting replications that appear to be endemic to the social science research infrastructure (e.g., Carpenter, 2012; Hartshome & Schachner, 2012; Makel, 2014; Makel & Plucker, 2014; Schmidt, 2009; Spellman, 2012).

For example, some posit that "successful replications are unpublishable; journals reject such research saying 'but we already knew that'" (Spellman, 2012, p. 58). Such systemic biases are well established and include the following:[2]

1. Submission bias. Conducting research and submitting for publication is time-consuming, and investigators may purposefully remove replications from the publication process to focus on other projects or because they believe replications cannot be published (e.g., Schlosberg, 1951; Spellman, 2012).
2. Funding bias. Research, including and especially RCTs, requires resources, making replications difficult to conduct if not funded (e.g., Schmidt, 2009).
3. Editor/reviewer bias. Journal editors and reviewers may be more likely to reject replications, driven by an implicit (or even explicit) belief that replications are not as prestigious as nonreplication articles (e.g., Makel, 2014; Neuliep & Crandall, 1990, 1993a, 1993b; Smith, as cited in Yong, 2012).
4. Journal publication policy bias. Journals may have policies against publishing replications (e.g., Madden et al., 1995; Ritchie et al., 2012; Smith, as cited in Yong, 2012).
5. Hiring bias. Institutions may not hire researchers who conduct replications, with Biases 2 and 3 possibly playing a role in these decisions.
6. Promotion bias. Similar to hiring bias, organizations may not value replication research as favorably as new and groundbreaking research within promotion and tenure activities (e.g., Madden et al., 1995).
7. Journals-analyzed bias. Previous research analyzing replication rates may have selected journals that publish few replications. Because each journal has its own editorial policies, it may be that some journals are more likely to accept replications than others (e.g., Ritchie et al., 2012).
8. Novelty equals creativity bias. Editors, reviewers, and researchers value creative contributions, but novelty and creativity are not synonymous. Most definitions of creativity and innovation propose criteria of novelty *and* utility; a novel result that cannot be replicated is by definition not creative (e.g., Makel & Plucker, 2014; Plucker, Beghetto, & Dow, 2004).

These biases may not uniformly deny publication of replications, but they certainly impede the process. Perhaps the most baffling aspect is that these biases exist even though the call for replications has existed for generations (e.g., Ahlgren, 1969; Bozarth & Roberts, 1972; Cohen, 1994; Rosenthal, 1969; Tukey, 1969).[3] Indeed, the incongruity between need and action has not gone unnoticed. Furchtgott (1984), in a discussion of the need to alter the outlook on publishing replications, stated that "not only will this have an impact on investigations that are undertaken, but it will reduce the space devoted to the repetitious pleas to replicate experiments" (p. 1316).

Replication in Other Scientific Domains

The concern over replication exists in many research domains, including advertising (Madden et al., 1995), biology (Begley & Ellis, 2012; Powell, 2007), economics (Anderson, Greene, McCullough, & Vinod, 2005; Dewald, Thursby, & Anderson, 1986; Kane, 1984), library sciences (Winters, 1996), marketing (Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007; Hubbard & Armstrong, 1994), medicine (Ioannidis, 2005a, 2005b, 2005c), political science (Golden, 1995; King, 1995), public health (Valentine et al., 2011), and sociology (La Sorte, 1972). A December 2011 special section in *Science* (Jasny, Chin, Chong, & Vignieri, 2011) discussed replication and its application in primate cognition (Tomasello & Call, 2011), field biology (Ryan, 2011), computer science (Peng, 2011), and genomics (Ioannidis & Khoury, 2011).

Using health care research as an example, Ioannidis (2005a), in a review of highly cited medical publications (i.e., those cited more than 1,000 times), found only 44% of replications produced results similar to the original study. Unsuccessful replications were most common when the original studies were not randomized and had small samples, both of which are common features of education research (especially compared to clinical medical research). In a separate study attempting to replicate highly cited cancer trial studies, researchers were able to successfully replicate only 6 of 53 trials, a success rate of just over 11% (Begley & Ellis, 2012). Similarly, researchers from the Bayer drug company were able to replicate only 35% of the published research findings they analyzed (Prinz, Schlange, & Asadullah, 2011). With such low replication success rates in domains known for methodological rigor and large samples sizes, the need for replication within the social sciences becomes more acute.

Replication has received the most attention of late in psychology (e.g., *Perspectives on Psychological Science* special issue, 2012; Yong, 2012). Recent conversations in psychological science over the publication of a controversial study on extrasensory perception (Bem, 2011) along with a few well-publicized cases of fraud (by eminent researchers) have energized discussion around what can be done to increase confidence in research findings. Moreover, the rate at which researchers accurately predict the outcomes of their own studies appears to support Bem's (2011) findings that precognition exists (e.g., Fanelli, 2010, 2012; Sterling, 1959; Sterling et al., 1995). Fanelli (2010) found that 91.5% of psychology studies supported the predicted outcome, making psychologists nearly 5 times better at predicting results than actual rocket scientists (i.e., space scientists). Moreover, the 91.5% success rate actually represents a decrease from the 97% success rate reported by Sterling (1959). Simmons and colleagues (2011) note that this hyperaccurate prediction record is

probably due to several factors, including collecting data until the desired result is found, not reporting unsuccessful trials, and eliminating observations and variables post hoc that do not support the targeted hypotheses. Nevertheless, according to a recent simulation study, publishing all studies—not just those that are statistically significant—leads to more accurate estimates of actual effects and differences (de Winter & Happee, 2013).

Data on replication rates are available for a few scientific fields. Makel, Plucker, and Hegarty (2012) analyzed the complete publication history of the 100 psychology journals with the highest 5-year impact factors and reported that only 1.07% of psychology publications were replications. Moreover, they noted the rate at which replications are being published is rising, albeit slowly (roughly 2% of publications since 2000 compared to less than 1% of publications in the 1980s and earlier). Contrary to the failure to replicate results in the medical field, less than 10% of psychology replications failed to replicate previous findings.

In a similar analysis of marketing research journals, Evanschitzky et al. (2007) analyzed nearly 1,400 articles from 1990 to 2004, and Hubbard and Armstrong (1994) analyzed the same journals from 1974 to 1989. Nearly 2.5% of the articles from 1974 to 1989 were replications, compared to 1.2% of articles from 1990 to 2004. This trend suggests a decrease in replication research, despite the fact that both studies found that the majority of those studies failed to replicate the original research!

Nevertheless, the concern about a dearth of replications is not universal. One journal editor claims, "I would wager a year's associate-editor pay that most [*Academy of Management Journal*] articles include at least partial replication, albeit not exact and, of course, not labeled 'replication research'" (Eden, 2002, p. 842). Similarly, an analysis of communication journals reported that 28% of studies had some form of replication, but only 3% clearly identified themselves as such (Kelly, Chase, & Tucker, 1979). This kind of masking is expected when the contents of rejection letters say things like a replication "translates into a minimal contribution to the field" (Sterling et al., 1995, p. 109). Similarly, 52% of surveyed social science editors reported that being a replication contributes to being rejected for publication. In fact, the only factors associated more strongly with rejection were the paper being published in the proceedings of a national (61%) or regional (53%) conference and an experiment that did not have a control group (54%; S. Kerr, Tolliver, & Petree, 1977). With such a high rejection rate, the disincentives to attempt replications are considerable. With the obvious lack of replicability in the medical studies discussed above, the concern over the veracity of some bedrock empirical beliefs should be high, making a lack of published replications a major weakness in any empirical field.

Replications in Education Research

The first use of the term *replication* in an education journal appeared in a 1938 paper in the *Journal of Educational Research* titled "An Example of Replication of an Experiment for Increased Reliability" (C. Peters, 1938). Focusing on the importance of relying on more than one implementation of an experiment, C. Peters (1938) emphasized the importance of relying on independent tests to understand the reliability of a particular finding (e.g., does one teaching method lead to better performance than another?). Given the current conversation regarding the importance of replication, the paper closes with great prescience,

> It is best not to place much confidence in a mathematically inferred ratio as far as its exact size is concerned but to stop with the assurance that a set of differences prevailing in the same direction indicates greater reliability than that expressed by the ratios of the samples taken singly. (C. Peters, 1938, p. 9)

Like many of the domains listed above, education research has several examples of notable replications, including recent research on paying students for performance (e.g., Fryer, 2011; Levitt, List, Neckermann, & Sadoff, 2011) as well as the impact of merit pay on teachers (e.g., Fryer, 2013; Yuan et al., 2013). Numerous studies have been conducted on each of these ideas and are not considered redundant or lacking in value.[4] However, to our knowledge, there have been no systematic investigations of replication in educational research. The current study applies the replication lens to education research by providing an overview of replications rates in education research journals. If the biases against replications in other fields extend to educational research, one would expect that replications in education would be extremely rare. Three broad sets of questions drove our investigation. First, how many replications are being published, and is the number of published replications changing over time? Second, what types of replications are being conducted; are direct or conceptual replications being conducted, and are they being conducted by the authors who conducted the original research or by a unique team? Finally, we investigated the extent to which the original findings were successfully replicated.

Method

The top 100 education journals (all types) according to 5-year impact factors were gathered using the online search engine ISI Web of Knowledge Journal Citation Reports, Social Sciences Edition (2011). In January 2013, using Web of Knowledge, the entire publication history of each of these 100 journals was searched to identify the total number of articles published as well as the number of articles that contained the search term *replicat** in the text. This method is similar to what has previously been used when searching the publication histories of large databases (e.g., Fanelli, 2010, 2012; Makel et al., 2012).

To estimate the actual replication rate (i.e., the percentage of articles that are replications), all of the articles that used the term *replicat** were analyzed. This analysis assessed (a) whether the term was used in the context of a new replication being conducted (as opposed to referring to gene replication) and, if so, (b) whether it was a direct or conceptual replication, (c) whether the replication was considered a success or a failure (success meaning the replicating authors conclude that their findings are similar to, or in the same direction as, the original findings), (d) whether the replicated article was written by the same authors (defined as having an overlap of at least one author), and (e) whether it was published in the same journal. The number of times the replicating and replicated articles have been cited were also recorded in April 2013 (if multiple studies were being

replicated, the average citation count of the replicated studies was calculated; the citation counts of books were not recorded because they are not calculated by Web of Knowledge).

All of the data were collected by the first author. The second author was given a set of written instructions (similar to the paragraphs above) to score a randomly selected subset of articles. In 18 out of 20 cases, articles were coded similarly, with minor, resolvable differences on the two remaining papers. This process provided evidence that the method identifying replications is itself replicable. The articles using *replicat** were then split and independently coded by the authors.

Gain ratios are also reported to help communicate changes in replication rates. The gain ratio statistic is similar to an odds ratio, but rather than being based on odds, it is based on the probability of an outcome (Agresti, 2007). If the probability of two events is equal (e.g., flipping a coin and getting heads vs. getting tails), the gain ratio is 1.0 and is considered significantly different from 1.0 if its 95% confidence interval does not include 1.0.

Results

The average 5-year impact factor of the top 100 journals in education was 1.55 (range = 0.52 to 5.46). Overall, 461 out of 164,589 articles from these education journals contained the term *replicat**, with 18 journals never using the term. Of the articles containing *replicat**, 221 (47.9%) were actual replications, giving an overall replication rate of 0.13% (221 out of 164,589; see Table 1 for a breakdown by journal) for the field. As a comparison, the estimated replication rate in psychology (Makel et al., 2012) was eight times (95% CI [6.99, 9.17]) higher than the replication rate in education journals. Only six journals had a replication rate over 1%, and 43 journals published no replications. Within the current sample, there does not appear to be a relationship between 5-year impact factor rank and replication rate ($r = -.03$, $p = .795$), although it should be noted that several of the journals analyzed were review journals that typically do not publish articles featuring new data. There may be journals outside the top 100 5-year impact factor that are publishing replications.[5] However, if this is the case, we worry that replications are being relegated to such low-visibility outlets that their ability to impact subsequent work is severely stunted (i.e., the 5-year impact factor of the 100th ranked journal was 0.52).

The articles that used the term *replicat** but were not actual replications typically used the term in the context of stating that the results needed to be replicated or in terms of replicating lesson plans. It should be noted that 12.7% of replications were replicating a finding from within the same (usually multistudy) article. Although not lacking value, within-article replications do not combat potential experimenter bias, error, or fraud.

As shown in Table 2, of the articles that were determined to be actual replications, 69.2% were conceptual, 28.5% were direct, and 2.3% included facets of both (i.e., usually in multistudy papers). Regarding outcomes, 67.4% of the replications reported successfully replicating the findings of the original study, 19.5% had mixed findings (supporting some, but not all, findings), and 13.1% failed to replicate the original findings. Interestingly, comparison of success rates by type of replication revealed that 71.4% of direct replications were successful compared to 66% of conceptual replications, with direct replications trending more successful but not significantly so, $\chi^2(4) = 5.95$, $p = .203$, Cramer's $V = .12$.

Only 30.6% of replications of previously published research were published in the same journal as the original study (replications from within the same article were not included for this calculation). But more interestingly, nearly half (48.2%) of the replications were conducted by the same research team that published the original research. The success rates of replications were significantly different based on whether there was author overlap; when replications were in the same publication as the original findings, 88.7% of replications were successful. When replications were in a new publication, but at least one author was on both the original and replicating articles, 70.6% of replications were successful. However, when there was no author overlap, only 54% of replications were successful, $\chi^2(4) = 21.03$, $p < .001$, Cramer's $V = .22$. Although same-author replications certainly contribute to research knowledge, this type of replication may not account for potential experimenter bias (regardless of whether the bias is intentional or unintentional). It is also worth noting that the recent, high-profile fraud cases within psychology often involved researchers replicating their own fraudulent studies with fraudulent replication data.

As can been seen in Figure 1, the rate at which replications are being conducted has increased in the last few decades. Since 1990, the replication rate has been 3.92 times higher (95% CI [2.75, 5.58]) than in previous years. Put another way, replications have gone from being 1 out of every 2,000 education articles to being roughly 1 out of every 500.

The median citation count of the replication articles was 5 (range = 0 to 135), whereas the median for the articles being replicated was 31 (range = 1 to 7,644) times. The original articles have had more time to be cited because they are older than their replicating counterparts (median publication year of 1991 and 1998, respectively[6]), but 5 citations are hardly insignificant given that only 1 of the top 100 education journals has a 5-year impact factor higher than 5.

Discussion

The present study analyzed the publication histories of the education journals with the top 100 five-year impact factors and found 0.13% of education publications were replications, substantially lower than the replication rates of previously analyzed domains. Contrary to previous findings in medical fields (e.g., Begley & Ellis, 2012; Ioannidis, 2005c), but similar to psychology research (Makel et al., 2012), the majority (67.4%) of education replications successfully replicated the original studies. However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles. This difference raises questions regarding potential biases in replicating one's own work and may be related to previous findings of questionable research practices in the social sciences (John et al., 2012; Simmons et al., 2011). More optimistically, same-author replications could merely be benefiting from the wisdom/experience from having done the study previously and thus may be able to more closely replicate the original methods. This phenomenon needs additional investigation.

Table 1
Replication Rates from the Top 100 Journals in Education Research

Journal Title	5-Year Impact Factor	Rank 5-Year Impact Factor	Articles Published	No. of Times *replicat** Appears	Replications Conducted	Replication Rate
Academic Psychiatry	1.05	64	1,296	2	0	0%
Academy of Management Learning & Education	4.05	2	693	0	0	0%
Adult Education Quarterly	0.62	94	901	3	3	0.33%
Advances in Health Sciences Education	2.06	22	635	7	3	0.47%
AIDS Education and Prevention	2.21	20	1,305	18	4	0.31%
American Educational Research Journal	3.09	5	1,953	13	10	0.51%
American Journal of Education	1.16	59	1,117	1	0	0%
Anthropology & Education Quarterly	0.74	85	1,311	1	0	0%
Applied Measurement in Education	0.85	81	373	6	2	0.54%
Australian Educational Researcher	0.52	100	312	0	0	0%
British Educational Research Journal	1.56	40	1,039	4	1	0.10%
British Journal of Educational Studies	1.17	57	2,869	2	0	0%
British Journal of Educational Technology	1.91	28	3,085	4	2	0.06%
British Journal of Sociology of Education	1.17	56	1,459	1	0	0%
Comparative Education	0.86	80	1,859	0	0	0%
Comparative Education Review	1.04	65	2,870	0	0	0%
Computers & Education	2.97	8	3,070	10	3	0.10%
Curriculum Inquiry	0.59	95	1,141	2	0	0%
Early Childhood Research Quarterly	2.61	11	678	13	7	1.03%
Economics of Education Review	1.44	45	1,266	2	1	0.08%
Education and Training in Developmental Disabilities	1.21	53	270	2	0	0%
Education and Urban Society	0.54	99	1,336	0	0	0%
Educational Administration Quarterly	1.39	48	1,390	5	3	0.22%
Educational Assessment Evaluation and Accountability	0.69	89	23	0	0	0%
Educational Evaluation and Policy Analysis	1.81	31	463	4	1	0.22%
Educational Gerontology	0.55	97	2,736	13	8	0.29%
Educational Policy	0.68	91	871	2	0	0%
Educational Research	0.93	74	355	3	2	0.56%
Educational Review	0.99	69	3,360	3	1	0.03%
Educational Studies	0.64	93	1,356	2	1	0.07%
Educational Technology Research and Development	1.65	38	7,280	9	4	0.05%
Elementary School Journal	1.51	41	963	4	4	0.42%
European Physical Education Review	0.77	84	161	1	0	0%
Foreign Language Annals	0.68	92	2,021	7	5	0.25%
Gender and Education	0.90	77	815	1	0	0%
Health Education Research	2.57	14	1,653	17	4	0.24%
Higher Education	1.31	51	3,081	3	2	0.06%
IEEE Transactions on Learning Technologies	0.93	73	166	1	0	0%
Innovations in Education and Teaching International	1.01	66	442	3	2	0.45%
Instructional Science	1.96	24	1,015	7	6	0.59%
Interactive Learning Environments	1.17	58	216	1	0	0%
International Journal of Computer-Supported Collaborative Learning	3.00	7	174	1	0	0%
International Journal of Educational Development	0.93	71	1,482	3	1	0.07%
International Journal of Science Education	1.72	35	2,052	10	6	0.29%
Journal of Adolescent & Adult Literacy	0.72	86	2,512	1	0	0%
Journal of American College Health	2.29	18	1,729	10	4	0.23%
Journal of College Student Development	1.18	55	4,483	10	7	0.16%
Journal of Computer Assisted Learning	1.76	33	771	7	2	0.26%
Journal of Curriculum Studies	0.97	70	2,292	0	0	0%
Journal of Diversity in Higher Education	0.87	79	18	2	1	5.56%
Journal of Education Policy	1.23	52	786	2	0	0%
Journal of Educational and Behavioral Statistics	2.44	16	464	2	0	0%
Journal of Educational Research	1.49	44	7,758	23	17	0.22%

(continued)

Table 1 (continued)

Journal Title	5-Year Impact Factor	Rank 5-Year Impact Factor	Articles Published	No. of Times *replicat** Appears	Replications Conducted	Replication Rate
Journal of Engineering Education	2.02	23	728	0	0	0%
Journal of Experimental Education	1.64	39	2,718	22	10	0.37%
Journal of Geography in Higher Education	1.42	46	1,141	2	0	0%
Journal of Higher Education	1.79	32	7,749	4	3	0.04%
Journal of Literacy Research	1.09	63	367	3	2	0.54%
Journal of Moral Education	0.72	87	1,715	4	2	0.12%
Journal of Philosophy of Education	0.56	96	819	0	0	0%
Journal of Research in Reading	1.50	43	294	7	4	1.36%
Journal of Research in Science Teaching	2.92	9	2,327	14	8	0.34%
Journal of School Health	1.91	27	5,784	11	4	0.07%
Journal of Social Work Education	1.11	61	1,711	9	3	0.18%
Journal of Teacher Education	2.23	19	3,903	1	0	0%
Journal of Teaching in Physical Education	1.41	47	678	0	0	0%
Journal of the Learning Sciences	3.08	6	347	3	1	0.29%
Language Learning	1.83	29	1,393	13	6	0.43%
Language Learning & Technology	2.47	15	288	0	0	0%
Language Teaching Research	0.91	76	243	1	1	0.41%
Learning and Instruction	3.73	3	620	16	8	1.29%
Minerva	1.00	68	1,682	0	0	0%
Oxford Review of Education	0.92	75	1,270	5	1	0.08%
Physical Review Special Topics-Physics Education Research	1.34	50	205	3	2	0.98%
Quest	0.69	90	830	1	0	0%
Reading and Writing	1.95	25	669	9	7	1.05%
Reading Research Quarterly	2.57	13	1,231	5	4	0.32%
Reading Teacher	0.88	78	11,006	1	1	0.01%
Research in Higher Education	1.83	30	1,324	8	7	0.53%
Research in Science Education	1.69	36	360	1	0	0%
Research in the Teaching of English	0.85	83	754	4	0	0%
Review of Educational Research	5.46	1	2,605	2	0	0%
Review of Higher Education	2.07	21	1,033	0	0	0%
Review of Research in Education	2.58	12	186	0	0	0%
School Effectiveness and School Improvement	1.16	60	424	4	3	0.71%
Science Education	2.32	17	1,551	6	1	0.06%
Scientific Studies of Reading	3.58	4	205	6	5	2.44%
Second Language Research	1.93	26	195	1	1	0.51%
Sociology of Education	2.73	10	1,639	12	7	0.43%
Sport Education and Society	1.10	62	372	0	0	0%
Studies in Higher Education	1.75	34	1,848	4	2	0.11%
Teachers College Record	1.19	54	7,846	3	0	0%
Teaching and Teacher Education	1.68	37	1,956	3	2	0.10%
Teaching in Higher Education	0.93	72	504	0	0	0%
Teaching of Psychology	0.55	98	2,827	18	7	0.25%
Tesol Quarterly	1.35	49	2,605	6	2	0.08%
Theory Into Practice	0.72	88	728	1	0	0%
Urban Education	1.01	67	1,525	0	0	0%
Vocations and Learning	1.51	41	82	0	0	0%
Zeitschrift fur Erziehungswissenschaft	0.85	81	576	1	0	0%

Note. Journal 5-year impact factors (based on citations in 2011 of articles published 2006 to 2010) and rankings are based on the 2011 Thomson Reuters ISI Web of Knowledge. Replication rate is the number of replications conducted divided by the total number of articles for that journal.

Table 2
Replication Results in Education Journals

	Replication Outcome			
Replication Type	**Success**	**Failure**	**Mixed**	**Total**
Direct	45 (71.4%)	11 (17.5%)	7 (11.1%)	63 (28.5%)
Conceptual	101 (66%)	18 (11.8%)	34 (22.2%)	153 (69.2%)
Mixed	3 (60%)	0 (0%)	2 (40%)	5 (2.3%)
Total	149 (67.4%)	29 (13.1%)	43 (19.5%)	221 (100%)
Replication Authorship	**Success**	**Failure**	**Mixed**	**Total**
Same Publication as Original	47 (88.7%)	2 (3.8%)	4 (7.5%9)	53 (24%)
Same Authors New Publication	48 (70.6%)	6 (8.8%)	14 (20.6%)	68 (30.8%)
All Unique Authors	54 (54%)	21 (21%)	25 (25%)	100 (45.2%)

Note. A total of 164,589 articles were analyzed using the complete publication history of the 100 journals with highest 5-year impact factors in 2011. Of the 461 that used the term *replicat** in text, 221 were deemed replications.

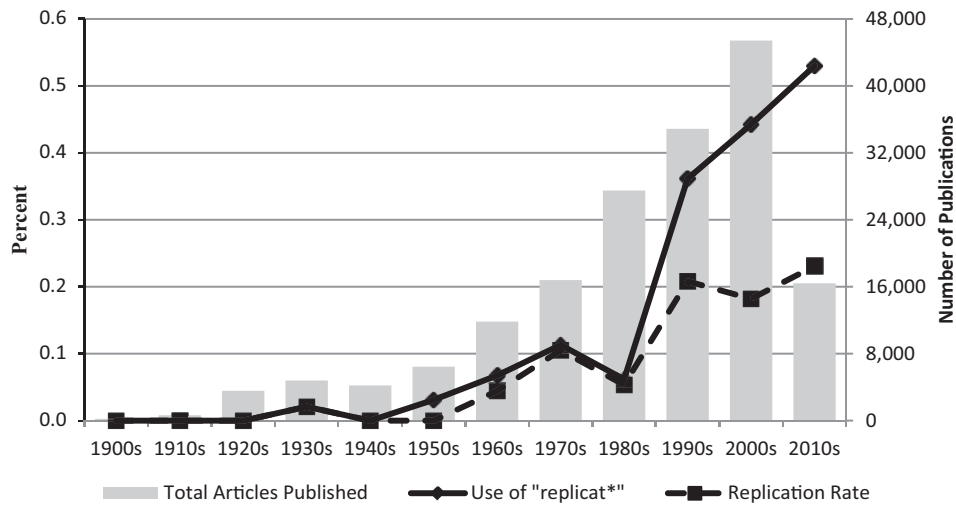

FIGURE 1. *Replication rate in education journals*
The solid line depicts the percentage of articles using replicat*. *The dotted line depicts the actual replication rate. Each bar represents the number of articles published in each decade. The 2010s data are based on data from only 2010 to 2012.*

Given such low replication rates, the need to increase is apparent and permeates all levels of education research. We believe that any finding should be directly replicated before being put in the WWC. We cannot know with sufficient confidence that an intervention works or that an effect exists until it has been directly replicated, preferably by independent researchers.

Thankfully, there is an abundance of proposed solutions to the dearth of replications. To help establish such confidence, the education sciences could emulate the plan recently announced in the journal *Perspectives on Psychological Science* that emphasizes the importance of "robust, replicable, and generalizable" research. This plan proposes a new article type, registered replication reports, which will "consist of multi-lab, high-quality replications of important psychology experiments along with comments by the authors of the original studies" (http://www.psychologicalscience.org/index.php/replication). The replication protocol is registered ahead of time, and independent labs all working from this protocol will be part of the eventual publication, with results reported in aggregate as well as by lab (see also Simons & Holcombe, 2014). Projects such as this would help bolster both credibility and understanding of published research. Although not part of this plan, one such "many-labs replication" has already been conducted (R. Klein et al., 2013), successfully replicating 10 of 13 attempted psychology studies.

Such initiatives will also help address the rampant problem of underpowered studies in the social sciences that allow large, but imprecise, effects sizes to be reported (e.g., Ioannidis, 2012; Johnson, 2013; McBee & Matthews, 2014a; Pashler & Harris, 2012; Schimmack, 2012). By fostering a system of preregistered replications, the focus of study can move away from achieving statistical significance and toward advancing precision in results (and, more to the point, advancing the education sciences). This is particularly true if replications focus on studies whose results may substantially change the outlook of the field, draw heightened attention (citations or media coverage), and/or have major policy implications.

Although universal standards of conducting replications have (obviously) not yet been adopted, some have been proposed (e.g., Brandt et al., 2014; Open Science Collaboration, 2012). Others have also rightly noted that training in conducting replications is also needed and can be implemented in graduate and undergraduate student projects (e.g., Frank & Saxe, 2012; Grahe et al., 2012). Paradoxically, effective training cannot begin until best practices are established. Similarly, no formalized method currently exists for how to compare findings of the original and replicating articles (i.e., no chi-square analysis or Bayesian priors are required or even typically used).[7] The norm is merely to report whether *p* values indicate whether direction of results is the same across studies. The newly proposed aggregating methods of registered replication reports of reporting independent and aggregated effect sizes will help compare original and replicated results as well as better estimate true effect sizes. Such methods slightly shift the emphasis away from the assessment of whether or not the replication succeeded in replicating the previous findings and toward winnowing the deep truth about the magnitude of effects. Similarly, determining what areas merit the resources needed to provide precise answers is an important question that currently has no definitive answer. Should it be left to individual researchers? Should editors request replications of specific studies? Should the major funding agencies devote resources specifically to the conduct of replications? The answer may be a combination of all of the above. But until the biases discussed in the introduction are removed, the point is moot; replication will not be conducted or published.

Other scholars have gone so far as to propose that findings be replicated prior to publication (e.g., Loevinger, 1968; Lubin, 1957; Neuliep & Crandall, 1993b; Roediger, 2012). This recommendation makes some sense, but it may not be practical—or possible—in many contexts (Lykken, 1968). Others have suggested the creation of a journal dedicated specifically to publishing replication studies (Ahlgren, 1969; Simons, 2012; Williams, 2012).

Specifically, Neuliep and Crandall (1993b) suggested that journals reserve a portion of their page space for replication research (see also Hubbard & Vetter, 1996), which has been implemented in other domains. For example, the *American Journal of Political Science* devoted a special section to publishing replications in 1996 and 1997 (Meier, 1995a). This policy resulted in a sharp increase in the number of published replications, an increase that quickly dissipated when the policy was discontinued. If page space is a concern, replication publications could be as short as a paragraph in length (Schlosberg, 1951). This could be particularly applicable for direct replications. Indeed, examples of such short-report replications exist but are rare (e.g., E. Klein, Gould, & Corey, 1969; Tarnowski, Drabman, Anderson, & Kelly, 1992).

Revising journal editorial policies to provide explicit encouragement for submitting replication studies—and reinforcing the importance of such studies with reviewers—would help ensure that the positive trend toward replications found in the current study continues. A few editors have already begun to do so (e.g., Eich, 2014; McBee & Matthews, 2014b). The increase in open, online journals may further encourage the submission of replication studies, as page limits essentially become moot in that context. Explicitly encouraging the submission of replications may also result in authors framing their submitted studies as replications.

All this being said, one replication, successful or failed, should neither cement nor condemn the original finding. The more replications (and the sooner they are conducted), the better. Replications will help uncover the precision with which we know size of the effects, not to mention the extent to which they generalize across contexts. As a field, we need to weed out false and narrow findings and buttress findings that generalize across contexts. In confirmatory research, preregistration of predictions, sample size, power needs, and so on could help avoid questionable research practices, such as data peeking and data hacking. We need to foster an environment in which being wrong is not a death knell and "massaging" data to avoid being wrong is. Research is a means to an end, and facts are more important to novelty.

Related Recent Initiatives

Many organizations and journals have recently announced changes relevant to the replication conversation (for longer overviews of such initiatives, see Makel, 2014; Makel & Plucker, 2014). For example, starting in January 2014, the journal *Psychological Science* requires all submissions to disclose facts about all data points excluded from analyses; all manipulations reported and not reported; all measures used in the study, including those not reported; and how sample size was determined (http://www.psychologicalscience.org/index.php/publications/observer/obsonline/whats-new-at-psychological-science.html). Additionally, the journal is attempting to promote open research practices, like data sharing, materials sharing, and preregistering design and analysis plans prior to data collection by adopting a badge system developed by the Open Science Collaboration (https://opensciceframework.org/project/TVyXZ/wiki/home/). Third, the journal is seeking to help researchers shift from a focus on *p* values toward a focus on "new statistics" (Cumming, 2014), such as effect sizes and confidence intervals. Similarly, the *PLoS* journals require all authors to post all relevant data and metadata publically (http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/). Open data and methods are a related and quite relevant topic in that they help the research community understand how original results were obtained as well as helping replicators design and conduct replication research while also providing a barrier to questionable research practices (e.g., John et al., 2012; Simmons et al., 2011).

A more systematic attempt at direct replication already underway is the Reproducibility Project (for a review, see Carpenter, 2012). By attempting to directly replicate findings from all the 2008 issues of *Journal of Personality and Social Psychology; Psychological Science*; and *Journal of Experimental Psychology: Learning, Memory, and Cognition*, this group is seeking to uncover common obstacles in conducting replications and predictors of replication success. To accomplish this, the project has teams of researchers following a set protocol to conduct replications with sufficient power.

Not all have been in full support of increased replication work. Although writing cautions against putting too much emphasis on and trust in replication attempts, some (e.g., Bissell, 2013; Cesario, 2014) have proposed that published findings

should be treated with deference until shown otherwise and that false findings will eventually be discovered and discarded. This seems like a catch-22; without replication attempts, such weeding out of false findings will be dramatically prolonged.

If a finding does not replicate due to particular circumstances (e.g., the quality of the teacher or students, the demographics of the sample, or the classroom climate), then that substantially weakens the original finding, not the replication. If an effect is so narrow/fragile that it can be found only in certain environments under certain circumstances (or by certain researchers), such limitations either need to be articulated in the original study (e.g., the effects reported in this study are limited to only environments in which the original authors are collecting the data) or need to be uncovered to avoid overgeneralizing the implications of the original research. If such boundary conditions are not reported by the original authors, replication attempts are needed to identify them.

Finally, we cannot ignore the rash of research fraud allegations in the social sciences in recent years, many of which allegedly went undetected for extended periods of time and a few of which have been confirmed beyond a reasonable doubt. The aftermath of just these few cases has been (a) policymaker and public distrust of research in certain fields, (b) seriously stained careers for the former students and colleagues of the perpetrators, and (c) legal, ethical, and reputational headaches (to put it mildly) for the institutions where the fraud occurred. The education sciences have been largely free of such scandals in recent years, but the odds are long that not a single such case exists within the field. Although retraction of questionable studies is often used to address this problem, retractions of research papers are often ignored or can take years to take hold (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). In the absence of a culture of replication, obvious deterrents are lessened, and any fraud will likely be discovered only after an extended period of time.

A case in point is the 1998 study purporting a link between common childhood vaccinations and development of autism.[8] The impact of the study was tremendous, causing significant public debate about the safety of vaccines, with many sources crediting the study for a still-growing antivaccination movement that has been indirectly linked to a resurgence of potentially deadly—and once nearly eradicated—diseases. Studies refuting the vaccine-autism link, yet using very different methodologies, were published almost immediately (e.g., Taylor et al., 1999), but the best-known direct replication, which also failed, was published only a full decade later (Hornig et al., 2008). After a series of investigations found evidence of widespread misconduct in the original study, the journal retracted it (Editors of *The Lancet*, 2010). Science may be self-correcting, but the often glacial pace of that correction does not match the speed of dissemination when results enter the public consciousness. Would some of the remedies suggested above, such as requiring replication of the study before publication, have prevented this situation? Perhaps, or perhaps not, but given the tremendous negative impact of the study, it is difficult to argue that the situation could have possibly been worse.

The Lancet's retraction case may be an outlier, but the number of such outliers is rapidly on the rise. In the first decade of the 21st century, the number of academic articles within Web of Science increased 44%, but the number of retractions increased over 1,300% from 30 per year to over 400 (Van Noorden, 2011). The growth in retractions illustrates the need for making use of the arsenal of tools at our disposal for deciphering fact from fiction. Directly replicating the results of others is a vital part of that process.

Conclusions

Like Campbell and Stanley (1963) noted a half century ago about experimental design, replication is not a panacea (Makel et al., 2012; Pashler & Wagenmakers, 2012). It will not resolve all issues and concerns about rigor, reliability, precision, and validity of education research. However, implicitly or explicitly dismissing replication indicates a value of novelty over truth (Nosek, Spies, & Motyl, 2012) and a serious misunderstanding of both science and creativity. If education research is to be relied upon to develop sound policy and practice, then conducting replications on important findings is essential to moving toward a more reliable and trustworthy understanding of educational environments. Although potentially beneficial for the individual researcher, an overreliance on large effects from single studies drastically weakens the field as well as the likelihood of effective, evidence-based policy. By helping, as Carl Sagan (1997) noted, winnow deep truths from deep nonsense, direct replication of important educational findings will lead to stronger policy recommendations while also making such recommendations more likely to improve education practice and, ultimately, the lives of children.

NOTES

[1]See http://www.bestevidence.org/aboutbee.htm.

[2]Crocker and Cooper (2011) also point out that an academic culture that reviles replication makes uncovering fraudulent research extremely difficult and extends the length of time fraudulent (and all false) findings stand unquestioned.

[3]Although it should be noted that the view of too few replications is not universal (cf. Bissell, 2013; Ostrom, 1971).

[4]The American Educational Research Association (AERA) explicitly addresses the importance of replication and the responsibility of education researchers to present and preserve their work in a form to enable replication. In 2006, AERA was the first among research societies in the social and behavioral sciences to issue Standards for Reporting on Empirical Social Science Research in AERA Publications that note the importance of replication in Standards 3.2, 5.6, and 7.5. The AERA Code of Ethics adopted in 2011 also speaks to standards for reporting on research and data sharing to allow for replication. AERA is currently turning its attention to ways that AERA and the journal publications program can further support and encourage replication studies. The decision by AERA in 2013 to publish *AERA Open* as an open-access journal is one vehicle to encourage the publication of peer-reviewed replication studies (Felice Levine, AERA Executive Director, personal communication, December 2013).

[5]The 2011 JCR Social Science Edition of ISI Web of Knowledge does not have a calculated entry for 5-year impact factor for *Educational Researcher* (*ER*), and so it is not included in the current results. Regardless, the *replicat** search term indicates 0 results for ER, so inclusion of the journal would not appreciably change the reported results.

[6]Only articles that replicated previously published findings were included in this comparison; articles that replicated only another study from the same publication were not included.

[7]The authors appreciate an anonymous reviewer who noted this point.

[8]At the time of this writing, the original, redacted paper has been cited over 1,700 times. We would prefer not to add to its negative impact by citing it again, although the study can be found by accessing the cited retraction.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York, NY: Wiley.

Ahlgren, A. (1969). A modest proposal for encouraging replication. *American Psychologist*, *24*, 471. doi:10.1037/h0037798

Anderson, R. G., Green, W. H., McCullough, B. D., & Vinod, H. D. (2005). *The role of data and program code archives in the future of economic research*. Working paper, Federal Reserve Bank of St. Louis, MO. Retrieved from http://research.stlouisfed.org/wp/2005/2005–014.pdf

Bakan, D. (1966). Test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437. doi:10.1037/h0020412

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives in Psychological Science*, *7*, 543-554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678. doi:10.3758/s13428–011–0089–5

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

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi:10.1037/a0021524

Bissell, M. (2013). The risks of the replication drive. *Nature*, *503*, 333–334.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, *27*, 774–775. doi:10.1037/h0038034

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?, *Journal of Experimental Social Psychology*, *50*, 217–224.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1561.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*, 40–48.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral Brain Science*, *14*, 119–186.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037//0003–066x.49.12.997

Cole, S. (1992). *Making science: Between nature and society*. Cambridge, MA: Harvard University Press.

Collins, H. M. (1985). *Changing order*. London, UK: Sage.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science*, *334*(6060), 1182. doi:10.1126/science.1216775

Cumming, G. (2014). The new statistics: Why and how. *Perspectives on Psychological Science*, *25*, 7–29.

de Winter, J., & Happee, R. (2013). Why Selective publication of statistically significant results can be effective. *PLoS One*, *8*, e66463. doi:10.1371/journal.pone.0066463

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The journal-of-money-credit-and-banking project. *American Economic Review*, *76*, 587–603.

Eden, D. (2002). Replication, meta-analysis, scientific progress, and AMJ's publication policy. *Academy of Management Journal*, *45*, 841–846.

Editors of *The Lancet*. (2010). Retraction: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *375*(9713), 445.

Eich, R. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*, 411–415. doi:10.1016/j.jbusres.2006.12.003

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.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives in Psychological Science*, *7*, 600–604.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics*, *126*, 1755–1798.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, *31*, 273–407.

Furchtgott, E. (1984). Replicate, again and again. *American Psychologist*, *39*, 1315–1316. doi:10.1037/0003–066X.39.11.1315.b

Golden, M. A. (1995). Replication and non-quantitative research. *PS: Political Science & Politics*, *28*, 481–483. doi:10.2307/420313

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives in Psychological Science*, *7*, 605–607.

Hansen, W. B. (2011). Was Herodotus correct? *Prevention Science*, *12*, 118–120. doi:10.1007/s11121–011–0218–5

Hartshome, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, *6*, 1–14, doi:10.3389/fncom.2012.00008

Hornig, M., Briese, T., Buie, T., Bauman, M. L., Lauwers, G., Siemetzki, U., . . . Lipkin, W. I. (2008). Lack of association between measles virus vaccine and autism with enteropathy: A case-control study. *PLoS One*, *3*(9), e3140.

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*, 233–248. doi:10.1016/0167–8116(94)90003–5

Hubbard, R., & Vetter, D. E. (1996). An empirical comparison of published replication research in accounting economics, finance, management, and marketing. *Journal of Business Research*, *35*, 153–164.

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228. doi:10.1001/jama.294.2.218

Ioannidis, J. P. A. (2005b). Contradictions in highly cited clinical research: Reply. *Journal of the American Medical Association*, *294*, 2696–2696. doi:10.1001/jama.294.21.2696-a

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

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives in Psychological Science*, *7*, 645–654.

Ioannidis, J. P. A., & Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science*, *334*, 1230–1232. doi:10.1126/science.1211811

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Introduction: Again, and again, and again . . . *Science*, *334*, 1225–1225.

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

Johnson, V. (2013). Revised standard for statistical evidence. *Proceedings of the National Academy of Science*, *110*, 19313–19317.

Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors*, *52*, 586–595. doi:10.1177/0018720810384394

Kane, E. J. (1984). Why journal editors should encourage the replication of applied econometric research. *Quarterly Journal of Business and Economics*, *23*, 3–8.

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*, 338–342.

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

Kerr, S., Tolliver, J., & Petree, D. (1977). Manuscript characteristics which influence acceptance for management and social-science journals. *Academy of Management Journal*, *20*, 132–141. doi:10.2307/255467

King, G. (1995). Replication, replication. *Political Science & Politics*, *28*, 444–452. doi:10.2307/420301

Klein, E. B., Gould, L. J., & Corey, M. (1969). Social desirability in children: An extension and replication. *Journal of Consulting and Clinical Psychology*, *33*, 128. doi:10.1037/h0027339

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2013). Investigating variation in replicability: The "many labs" replication project. Retrieved from https://osf.io/wx7ck/files/ManyLabsManuscript.pdf/

Kruskal, W., & Majors, R. (1989). Concepts of relative importance in recent scientific literature. *American Statistician*, *43*, 2–6. doi:10.2307/2685157

La Sorte, M. A. (1972). Replication as a verification technique in survey research: A paradigm. *Sociological Quarterly*, *13*, 218–277. doi:10.1111/j.1533–8525.1972.tb00805.x

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379. doi:10.1037/a0025172

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2011). *The impact of short-term incentives on student performance*. Retrieved from http://bfi.uchicago.edu/events/20111028_experiments/papers/Levitt_List_Neckermann_Sadoff_Short-Term_Incentives_September2011.pdf

Levy, L. H. (1969). Reflections on replications and the experimenter bias effect. *Journal of Consulting and Clinical Psychology*, *33*, 15–17.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*, 106–131.

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *American Statistician*, *47*, 217–228.

Loevinger, J. (1968). The "information explosion." *American Psychologist*, *23*, 455. doi:10.1037/h0020800

Lubin, A. (1957). Replicability as a publication criterion. *American Psychologist*, *12*, 519–520. doi:10.1037/h0039746

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159. doi:10.1037/h0026141

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising*, *24*, 77–87.

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

Makel, M. C., & Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 27–29.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science*, *7*, 537–542.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

McBee, M. T., & Matthews, M. S. (2014a). Change starts with journal editors: In response to Makel (2014). *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 8–10.

McBee, M. T., & Matthews, M. S. (2014b). Welcoming quality non-significance and replication work, but not the *p*-values: Announcing new policies for quantitative research. *Journal of Advanced Academics*, *25*, 68–78.

Meier, K. J. (1995a). Publishing replications: OK, let's try it. *Political Science & Politics*, *28*, 662–663. doi:10.2307/420515

National Reading Panel. (1999). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, *5*, 85–90.

Neuliep, J. W., & Crandall, R. (1993a). Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior and Personality*, *8*, 1–8.

Neuliep, J. W., & Crandall, R. (1993b). Reviewer bias against replication research. *Journal of Social Behavior and Personality*, *8*, 21–29.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science*, *7*, 615–631.

Open Science Collaboration. (2012). An open, large-scale collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Ostrom, T. H. (1971). To replicate or explicate. *American Psychologist*, *26*, 312. doi:10.1037/h0020338

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives in Psychological Science*, *7*, 531–536.

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

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*, 1226–1227. doi:10.1126/science.1213847

Perspectives on Psychological Science. (2012). Special section on replicability in psychological science: A crisis of confidence? Retrieved from http://pps.sagepub.com/content/7/6.toc

Peters, C. C. (1938). An example of replication of an experiment for increased reliability. *Journal of Educational Research*, *32*, 3–9.

Peters, D. P., & Ceci, S. J. (1982). Peer-reviewed practices of psychological journals: The fate of accepted published articles, submitted again. *Behavioral Brain Science*, *5*, 187–195.

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*. Advance online publication. doi: 10.3102/0013189X13507104

Plucker, J., Beghetto, R. A., & Dow, G. (2004). Why isn't creativity more important to educational psychologists? Potential, pitfalls, and future directions in creativity research. *Educational Psychologist*, *39*, 83–96.

Powell, K. (2007). Going against the grain. *PLoS Biology*, *12*, 2748–2753.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.
Rhine, J. B. (1997). *Extra-sensory perception*. Boston, MA: Branden. (Original work published 1934)
Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Replication, replication, replication. *Psychologist*, *25*, 346–348.
Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *Academic Observer*. Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-12/psychologys-woes-and-a-partial-cure-the-value-of-replication.html
Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.
Rosenthal, R. (1967). Covert communication in the psychological experiment. *Psychological Bulletin*, *67*, 356–367. doi:10.1037/h0024529
Rosenthal, R. (1969). On not so replicated experiments and not so null results. *Journal of Consulting and Clinical Psychology*, *33*, 7–10. doi:10.1037/h0027231
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi:10.1037//0033–2909.86.3.638
Rotton, J., Foos, P. W., Vanmeek, L., & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. *Journal of Social Behavior and Personality*, *10*, 1–13.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. doi:10.1037/h0042040
Ryan, M. J. (2011). Replication in field biology: The case of the frog-eating bat. *Science*, *334*, 1229–1230. doi:10.1126/science.1214532
Sagan, C. (1997). *The demon-haunted world: Science as a candle in the dark*. New York, NY: Ballantine Books.
Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566.
Schlosberg, H. (1951). Repeating fundamental experiments. *American Psychologist*, *6*, 177–177. doi:10.1037/h0056148
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi:10.1037/a0015108
Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, 470, 437–437. doi:10.1038/470437a
Simons, D. J. (2012). The need for new incentives. *Psychologist*, *25*, 349.
Simons, D. J., & Holcombe, A. O. (2014). Registered replication reports: A new article type at *Perspectives on Psychological Science. APS Observer*, *27*(3), 16–17.
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
Spellman, B. A. (2012). Introduction to the special section: Data, data, everywhere . . . especially in my file drawer. *Perspectives on Psychological Science*, *7*, 58–59. doi:10.1177/1745691611432124
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
Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician*, *49*, 108–112. doi:10.2307/2684823
Tarnowski, K. J., Drabman, R. S., Anderson, D. F., & Kelly, P. A. (1990). Disproportionate referrals for child academic behavior problems: Replication and extension. *Journal of Consulting and Clinical Psychology*, *58*, 240–243. doi:10.1037//0022–006x.58.2.240
Taylor, B., Miller, E., Farrington, C., Petropoulos, M. C., Favot-Mayaud, I., Li, J., & Waight, P. A. (1999). Autism and measles, mumps, and rubella vaccine: No epidemiological evidence for a causal association. *The Lancet*, *353*(9169), 2026–2029.
Tomasello, M., & Call, J. (2011). Methodological challenges in the study of primate cognition. *Science*, *334*, 1227–1228. doi:10.1126/science.1213443
Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91. doi:10.1037/h0027108
Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Commentaries on "Replication in Prevention Science": A rejoinder. *Prevention Science*, *12*, 123–125. doi:10.1007/s11121–011–0220-y
Van Noorden, R. (2011). The trouble with retractions. *Nature*, *478*, 26–28.
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., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives in Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078
What Works Clearinghouse. (2013). Home page. Retrieved from http://ies.ed.gov/ncee/wwc/
Whitehurst, G. (2003). *The Institute of Education Sciences: New wine, new bottles. A presentation by IES director Grover (Russ) Whitehurst*. Retrieved from http://www2.ed.gov/rschstat/research/pubs/ies.html
Williams, S. N. (2012). Replication initiative: Prioritize publication. *Science*, *336*, 801–802.
Winters, B. A. (1996). But I don't have anything to write about. *Library Acquisitions–Practice and Theory*, *20*, 391–394. doi:10.1016/s0364–6408(96)00068–3
Yong, E. (2012). Bad copy. *Nature*, *485*, 298–300.
Yuan, K., Le, V., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis*, *35*, 3–22.

AUTHORS

MATTHEW C. MAKEL, PhD, is a gifted education research specialist at the Duke University Talent Identification Program, 300 Fuller Street, Durham, NC 27701; *mmakel@tip.duke.edu*. His research focuses on research methods and academic talent development.

JONATHAN A. PLUCKER, PhD, is the Raymond Neag Endowed Professor of Educational Leadership and professor of educational psychology in the Neag School of Education at the University of Connecticut, 2131 Hillside Road, Storrs, CT 06269; *jonathan.plucker@uconn.edu*. His research focuses on education policy, talent development, and program evaluation.

Manuscript received July 31, 2013
Revisions received October 14, 2013, April 10, 2014, and June 4, 2014
Accepted June 26, 2014