Please use the following citation when referencing this work:

McGill, R. J., Ward, T. J., & Canivez, G. L. (2018). On the evidential value of school psychology intervention research. *The School Psychologist, 72 (3),* 48-57. Retrieved from https://apadivision16.org/the-school-psychologist-tsp/

On the Evidential Value of School Psychology Intervention Research

Ryan J. McGill

Thomas J. Ward

William & Mary

Gary L. Canivez

Eastern Illinois University

Author notes

Correspondence concerning this article should be addressed to Ryan J. McGill, William & Mary School of Education, P.O. Box 8795, Williamsburg, VA 23188. E-Mail: rmcgill@wm.edu

Abstract

The present investigation examined the evidential value of intervention research published in school psychology journals from 2016-2017. Although the nine journals that were reviewed favored studies that have a low probability of wrongly rejecting the so-called null hypothesis (H_0), *p*-curve analyses did not yield evidence of selective reporting in the literature. Whereas the potential threat of publication bias was not supported, the prevalence rate of replication studies was relatively low. Nevertheless, the majority of published intervention studies in school psychology contained sufficient power to promote future replication of observed effects. It is believed that the present investigation will be instructive for broadening the replication conversation in the school psychology to avoid problems experienced in other psychological research areas.

Keywords: replication crisis, *p*-curve, evidence-based practice

On the Evidential Value of School Psychology Intervention Research

There has been a longstanding debate in the field of school psychology regarding the role of research and its utility and applicability for school-based practitioners. Nevertheless, American Psychological Association and National Association of School Psychologists training and practice standards specify that practitioners should demonstrate skills to apply research findings as a foundation for service delivery regardless of their level of training, and the responsibility to acquire and apply accurate knowledge about effective practices are considered to be an epistemic responsibility of the clinician (e.g., O'Donahue & Henderson, 1999). As stated by Lilienfeld et al. (2012), "all school psychologists, regardless of the setting in which they operate, need to develop and maintain a skill set that allows them to distinguish evidence-based from non-evidence based practices" (p. 8) and this notion is a foundational principle of the broader evidence-based practice movement.

It stands to reason that implementing this skill set requires practitioners to place a tremendous amount of faith in the accuracy of published research and the integrity of the very publication process itself. Unfortunately, over the past decade, serious questions have been raised about commonly accepted methodologies (e.g., allegiance to null hypothesis significance testing [NHST]) in scientific research and the reproducibility of many published findings in psychological science. For example, Ioannidis (2005) suggested that half of all published research findings are likely false due to the prevalence of underpowered studies and the use of questionable research practices (QRPs)[1] and the results of a highly influential study published in

[1] These are two separate but equally important confounds. The power of study is the probability that it will distinguish between a true effect and chance and is mediated by sample size and the strength of the observed effect. On the other hand, QRPs are a class of techniques in which

Science seemed to confirm this contention. In that study, researchers associated with the Open Science Collaboration (2015) attempted to replicate 100 experiments reported in articles published in three high-ranking psychology journals in 2008 and were able to obtain a replication rate of only 39%. These results have prompted many to conclude that psychology is in the midst of a replication crisis[2]. Though it should be noted that some catalyst scholars reject this notion (see Baumeister & Vohs, 2016) and subsequent debates on these issues have been acrimonious. Regardless of one's position on whether psychological findings are replicable, the principal takeaway from these debates is that we need to fundamentally change the way that we think about interpreting data and results.

Low Statistical Power and the Prevalence of QRPs

As noted by Nosek, Spies, and Motyl (2012), incentive structures in science prioritize novelty and a publication bias against research that reports null effects is well known. Sterling, Rosenbaum, and Weinkam (1995), examined the publication decisions for 11 major journals and found that that 94% of studies reporting statistical tests in psychology rejected the null hypothesis (H_0) casting doubt on the representativeness of those findings. That is, if consumers accept these results at face value, they must conclude that virtually all studies that are published in the professional literature are performed with high power and under conditions in which investigators have formulated true hypotheses. Accordingly, the concept of statistical power (the probability of rejecting H_0 when it is false) is critical for understanding the role that publication bias may play in the so-called replication crisis. In particular, troubles arise when one tries to

researchers artificially increase the likelihood of rejecting the null hypothesis (i.e., data snooping, dropping cases, obtaining additional measurements until significance is obtained, etc.).

[2] Concerns about reproducibility are not limited to psychology and are widespread across scientific disciplines.

interpret a significant result from a study with low power. As an example, suppose a researcher reports a statistically significant effect in an intervention study with power at .50. If the same study were repeated with different samples under the same conditions, that effect would be observed in only 50% of the investigations. If all of these studies were submitted and accepted for publication, a pattern of contradictory findings would emerge and school psychologists conversant with this literature would be less likely to regard the intervention as an empirically supported practice. As a result, surveys (e.g., Szucs & Ioaannidis, 2017) indicating that the median estimated power in psychological research is approximately .30 are sobering and suggest that a non-trivial proportion of published studies are likely overestimating true effects or the product of a Type I statistical error.

A file drawer problem can occur when the probability of publication becomes dependent on statistical significance. In this type of culture, negative results are selectively reported or in some cases discarded entirely resulting in "a remarkable string of successes for psychological theories in published articles" (Heene & Ferguson, 2017, p. 43). In response, researchers may resort to using a number of QRPs to increase their chances of attaining significant results that are more likely to be published. These include data snooping (data mining to uncover patterns in data that can be presented as statistically significant), hypothesizing after results are known (HARKing), and *p*-hacking (exploiting researcher degrees of freedom until a significant *p*-value is obtained). How prevalent are such practices? In a survey about their involvement in QRPs, the self-admission rate among 2,000 psychologists ranged from 27% to 40% across disciplines (John, Loewenstein, & Prelec, 2012).

According to Simmons, Nelson, and Simonsohn (2011), the ubiquity of these practices make it "unacceptably easy to publish 'statistically significant' evidence consistent with *any*

hypothesis" (p. 1359) resulting in an epistemological confound they termed *false positive psychology*. At the root of this dilemma is the fact that QRPs increase the maximum false-positive rate beyond conventional nominal levels (i.e., 5%). This is not a trivial statistical matter. Whereas practices that are universally regarded as unethical such as fabricating data increase Type I error by 100%, some estimates indicate that QRPs can increase the false-positive rate by up to 60% (Schimmack, 2012). Put simply, false positives are costly errors. Once published in the literature, they may be used by practitioners and researchers as evidence to support potentially ineffective practices.

Correcting for Selective Reporting

The prevalence of QRPs suggest that rather than discarding entire studies, researchers may merely eliminate (file) the subsets of analyses that produce negative findings. This selective reporting is particularly insidious because it upends assumptions about the number of failed attempts needed to produce a false-positive result and invalidates the traditional "fail-safe" calculations that are used to assess the file-drawer problem in meta-analyses. As a potential safeguard, Simonsohn, Nelson, and Simmons (2014) introduced the *p*-curve method for detecting effects associated with selective reporting. The purpose of the *p*-curve is to detect evidential value by distinguishing between sets of significant findings that are likely due to selective reporting. A *p*-curve is the distribution of statistically significant *p* values for a set of studies and the shape of the distribution helps to uncover selective reporting versus true effects. Interpreting results is fairly straight forward: right-skewed curves may indicate evidential value (i.e., findings that are likely replicable), flat curves indicate no evidential value, and left-skewed curves may indicate the presence of selective reporting in the literature. Although *p*-curve analyses are being

increasingly used by researchers to defend *and* raise concern about the quality of research evidence in allied fields, they have yet to be reported in the school psychology literature.

Purpose of the Current Investigation

Unfortunately, substantive discussions of the replication crisis in the school psychology literature have been limited save a recent commentary by Shaw and D'Intino (2017). Thus, the impact and prevalence of selective reporting in school psychology remains largely unknown. To remediate this gap in the literature, the goal of the present study was to examine the evidential value of intervention research published in nine school psychology journals over a two-year period (2016-2017) with a specific emphasis on the potential threat of publication bias using the *p*-curve method and estimating the replication rate of published research in the field. Examination of these separate, but equally important issues, is important because it can lead to over estimates of effects in the empirical literature. Although a recent article by Villarreal and colleagues (2017) examined the characteristics of intervention research in school psychology journals, the evidential value of the studies was not assessed. It is believed that the results from the present investigation will be instructive for generating a much needed discussion about the quality of research practices in school psychology.

Method

Data collection and analyses for the present study occurred in several steps. First, the archives of nine school psychology journals (*Contemporary School Psychology*, *International Journal of School and Educational Psychology*, *Journal of Applied School Psychology*, *Journal of School Psychology*, *Psychology in the Schools*, *School Psychology Forum*, *School Psychology International*, *School Psychology Quarterly*, and *School Psychology Review*) were searched for all articles published from 2016-2017. As a preliminary screening, the abstracts for the articles

(*N* = 689) were reviewed to identify appropriate intervention articles. We focused specifically on locating articles that systematically evaluated intervention outcomes. That is, survey research examining the preferences and prevalence of practices among practitioners, and studies focused on the process of implementation[3] were excluded from further consideration. Intervention articles were extracted and evaluated in more detail to determine if they met *a priori* inclusionary criteria for the current study. In order to be included in the analysis, a statistical test result had to be associated with a determinable research hypothesis. In accordance with best practice, studies were not included if they were (a) commentary or editorial articles, (b) literature reviews or research summaries, (c) meta-analyses (to prevent reporting duplicate effects), (d) non-empirical case studies, or (c) reported results not compatible or able to be transformed to be compatible for *p*-curve analyses (i.e., exact *p* values). Next, we subjected the statistical effects from individual studies to *p*-curve analysis using the *p*-curve app version 4.06 (http://www.p-curve.com/) to determine the evidential value of the studies as a whole. In a *p*-curve analysis, the *p* values from a set of studies are plotted along a curve and then statistically evaluated for potential bias using a binomial sign test. A right-side bias in a curve is considered to be evidence for the presence of a real effect (i.e., replicable) whereas a flat curve or left side bias suggests a questionable effect that *may* be an artifact of selective reporting and/or QRPs (Simonsohn et al., 2014). Supplementary tables (see Tables X.1-X.3) containing summary information for the statistical effects that were included in the present analyses and the studies that did not meet inclusionary criteria are available in an online supplement (https://osf.io/zf648/).

[3] Unless the purpose of the study was to evaluate the effect of an intervention designed to promote intervention implementation or integrity.

Results

Descriptive statistics for the initial article search are reported in Table 1. Of the 689 articles that were published across school psychology journals from 2016-2017, 27% (n = 189) were intervention articles where the evaluation of outcomes was a primary objective. Among these studies, 43% (n = 81) disclosed the result of a statistical test(s), among which, 94% reported one or more statistically significant outcomes. The articles were also inspected to estimate the replication rate of intervention research in school psychology. Studies were coded as a *replication* if replication was noted as an explicit goal of the research within the manuscript. The resulting replication rate among the school psychology journals that were reviewed (~6%) over this time period is relatively consistent with published estimates in other fields (e.g. Makel, Plucker, & Hegarty, 2012). Of the journals examined in the present study, *School Psychology Review* was the only journal that posted a replication policy on its website. That policy statement indicated that replication studies would be considered for publication as a part of a special section of the journal.

Table 2 reports the results of *p*-curve analyses across the nine journals. Not surprisingly, the power estimates and percentage of statistically significant effects indicating evidential value (i.e., $p < .025$) varied significantly across the journals. Nevertheless, the Z-test for each *p*-curve was statistically significant indicating evidential value. The results of the omnibus *p*-curve analysis across journals is presented graphically in Figure 1[4]. Among the 242 total effects that were extracted from 71 different intervention studies (M = 3.40 effects per study), 160 were statistically significant (i.e., $p < .05$) and 122 (76%) were indicative of evidential value (i.e., $p < .025$). Visual inspection of the graph in Figure 1 reveals the desired right side bias in the curve

[4] Independent *p*-curve graphs for each journal are provided in the online supplement.

resulting in a statistically significant binomial sign test ($p < .05$, one-tailed) indicating that effects associated with the present set of studies are not likely the result of selective reporting in the literature. To wit, the estimated power associated with the statistically significant effects included in the *p*-curve is .81 (90% CI [.74, .86]).

Discussion

Due to a host of high-profile failures to replicate studies in social and experimental psychology, methodologists are in the early stages of examining the credibility of traditional scientific practices in the discipline. Although we contend that school psychology has much to learn from these conversations, the field remains insulated from on-going efforts to improve the state of psychological science (Tackett et al., 2017). In an effort to broaden the replicability conversation, the present study utilized the *p*-curve method to examine the evidential quality of intervention research published in several school psychology journals in order to determine the degree to which statistically significant findings reflected selective reporting rather than true effects. To our knowledge, this is the first application of *p*-curve analyses reported in the school psychology literature.

The present results are virtually identical to estimates furnished previously by Sterling, Rosenbaum, and Weinkam (1995). We found that the publication decisions in nine peer reviewed school psychology journals appear to disproportionality favor studies that observe effects that have a low value of incorrectly rejecting the null hypothesis. Of the studies that disclosed the results of a statistical test(s), 94% concurrently reported the results of at least one statistically significant finding. Although these results would seem to implicate the presence of publication bias, this hypothesis was not supported by results of the *p*-curve analyses.

With regard to the issue of replication, our examination of 189 intervention articles across the school psychology journals indicated a relatively low percentage of replication studies. While this rate is not considerably different from other disciplines, it suggests that published intervention findings in school psychology are rarely subjected to systematic replication. Although we stipulate that the operational definition employed in the current study is likely a conservative estimate of the actual replication rate given the fact that many intervention studies could be classified as *conceptual* replications of previous work, the fact that so few authors reported replication as being an explicit goal in the studies suggests that the rate of *direct* replications, which has been regarded by some as the cornerstone of science (e.g., Coyne, Cook, & Therrien, 2016), in our field is likely quite low.

Nevertheless, *p*-curve results indicated that the overwhelming majority of results reported in the intervention studies that were analyzed from 2016-2017 were of evidential value and the estimated power in the overall sample (.81) far surpasses median estimates that have previously been reported in the literature (e.g., Suzcs & Ionnidis, 2017). As a result, it is unlikely that these results are artifacts of selective reporting. Despite these positive findings, it is important to note that the *p*-curve method focuses only on the effects of selective reporting in the literature and is not useful for identifying other important QRPs such as HARKing, which may be of greater concern to the field given the fact that school psychologists frequently have access to large datasets and samples of participants when conducting studies. In contrast to *p*-hacking, HARKing and data snooping are almost impossible to identify absent study pre-registration as readers only see the final results in a published article and have no way of knowing how those results were *actually* produced (Schimmack, 2012). Pre-registration is usually accomplished by posting research plans in an independent registry prior to data collection so that consumers are

better able to distinguish exploratory from confirmatory research. Unfortunately, research pre-registration in school psychology is virtually nonexistent.

Study Limitations

In spite of these results, the *p*-curve method has several limitations. Most notably, it is not possible to include studies that do not report results produced from alternative to exact tests. Accordingly, many single-case designs and studies primarily reporting effect sizes are not able to be included in the online app at the present time. In the current study, 57 studies reporting intervention outcomes were unable to be included in the *p*-curve analyses because the statistical information necessary for extracting exact *p* values was not available. Of the aforementioned studies, 82% employed single-case design (SCD). Given the prevalence of SCD research in the school psychology literature, this limitation is particularly notable.

Additionally, the *p*-curve method is most often applied to investigate the quality of focal research programs and, in some cases, the results furnished by specific researchers and teams. Future investigations along these lines would be instructive. In doing so, it is important to keep in mind that selective reporting and other related QRPs are likely not the by-product of malicious intent and that they are a class of practices that are distinct from other behaviors such as data fabrication, which are clearly unethical (Nelson, Simmons, & Simonsohn, 2018).

Conclusion

The present study has substantive implications for school psychology research and practice. Given the recent high-profile replication failures in psychological science, efforts should be undertaken to encourage and promote a more robust culture of replication in the school psychology literature. Additionally, journal editors and reviewers can help to protect against the insidious effects of QRPs and selective reporting by giving equal consideration to high quality

studies that report non-significant results and encouraging authors to pre-register their study protocols in open source forums such as the Open Science Framework (Kratochwill, Levin, & Horner, 2018). On the other hand, the issues raised in the present article suggest that practitioners should guard against overinterpreting the results from isolated intervention studies without considering the broader literature associated with the application of that intervention (i.e., literature that may report negative or in some cases contraindicated effects) and the degree to which those effects have been replicated in the school psychology literature. Additionally, all school psychologists are encouraged to become conversant with the broader replication crisis literature in psychology as well as other allied fields (i.e., evidence-based medicine). We believe these efforts are crucial for advancing our science and furthering efforts to make school psychology incorruptible.

References

Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science, 11,* 574-575. doi: 10.1177/1745691616652878

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework for conceptual replications. *Remedial and Special Education, 37,* 244-253. doi: 10.1177/0741932516648463

Heene, M., & Ferguson, C. J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 34-52). West Sussex, UK: Wiley.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLos Medicine, 2,* e124. 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

Kratochwill, T. R., Levin, J. R., & Horner, R. H. (2018). Negative results: Conceptual and methodological dimensions in single-case intervention research. *Remedial and Special Education, 39,* 67-76. doi: 10.1177/0741932517741721

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing between science pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology, 50,* 7-36. doi: 10.1016/j.jsp.2011.09.006

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: 0.1177/1745691612460688

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69,* 511-534. doi: 10.1146/annurev- psych- 122216- 011836

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631. doi: 10.1177/1745691612459058

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

Shaw, S. R., & D'Intino, J. (2017). Evidence-based practice and the reproducibility crisis in psychology. *Communique, 45 (5),* 1-21. Retrieved from http://www.nasponline.org

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology, 143,* 534-547. doi: 10.1037/a0033242

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. *The American Statistician, 49,* 108-111. doi: 10.2307/2684823

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLos Biology, 3,* e2000797. doi: 10.1371/journal.pbio.2000797

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., …& Shrout, P. E. (2017). It's time

to broaden the replicability conversation: Thought for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742-756. doi: 10.1177/1745691617690042

O'Donohue, W., & Henderson, D. (1999). Epistemic and ethical duties in clinical decision-making. *Behaviour Change, 16,* 10-19. doi: 10.1375/bech.16.1.10

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 943. doi: 10.1126/science.aac4716

Villarreal, V., Castro, M. J., Umana, I., & Sullivan, J. R. (2017). Characteristics of intervention research in school psychology journals: 2010-2014. *Psychology in the Schools, 54,* 548-559. doi: 10.1002/pits.22012

Table 1

Outcome of Tests of Significance and Prevalence of Replication Studies for Intervention Research Published in Nine School Psychology Journals from 2016-2017

Journal	*k*	Intervention Articles [a]	Articles Reporting Statistical Tests	Statistically Significant [b]	Replication Studies [c]
CSP	74	24 (32%)	9 (38%)	8 (89%)	0 (0%)
IJSEP	64	16 (25%)	1 (6%)	1 (100%)	1 (6%)
JASP	38	15 (39%)	5 (33%)	4 (80%)	1 (7%)
JSP	84	23 (27%)	18 (78%)	16 (88%)	1 (4%)
PITS	166	43 (30%)	16 (37%)	14 (88%)	2 (5%)
SPF	51	19 (37%)	6 (32%)	5 (83%)	2 (11%)
SPI	76	14 (18%)	8 (57%)	8 (100%)	0 (0%)
SPQ	85	16 (19%)	7 (44%)	7 (100%)	3 (19%)
SPR	51	19 (37%)	11 (59%)	11 (100%)	2 (11%)
Total	**689**	**189 (27%)**	**81 (43%)**	**76 (94%)**	**12 (6%)**

Note. *k* = number of manuscripts published (2016-2017). CSP = Contemporary School Psychology; IJESP: International Journal of School and Educational Psychology; JASP = Journal of Applied School Psychology; JSP = Journal of School Psychology; PITS = Psychology in the Schools; School Psychology Forum; SPI = School Psychology International; SPQ = School Psychology Quarterly, SPR = School Psychology Review.

[a] Number of articles reviewed in the present study (i.e., intervention outcomes was a primary objective).

[b] Articles reporting a statistically significant (i.e., $p < .05$) outcome(s).

[c] Studies were coded as replication attempt if it was noted as an explicit goal of the research project.

Table 2

Results of P-Curve Analyses of Intervention Research Published in School Psychology Journals from 2016-2017

Journal	% Evidential Value [a]	Z-Test of Half *P*-Curve	Power Estimate [90% CI]	Evidential Value Based on *P*-Curve	Evidential Value Based on Power
1. CSP	56%	-2.71*	.46** [.07, .82]	Yes	Yes
2. IJSEP	Insufficient Data to Calculate Independent *P*-Curve				
3. JASP	64%	-4.79*	.64** [.31, .86]	Yes	Yes
4. JSP	69%	-5.72*	.55** [.34, .73]	Yes	Yes
5. PITS	83%	-3.49*	.55** [.34, .73]	Yes	Yes
6. SPF	87%	-5.25*	.91** [.73, .97]	Yes	Yes
7. SPI	83%	-9.44*	.92** [.82, .97]	Yes	Yes
8. SPQ	84%	-15.72*	.99** [.99, .99]	Yes	Yes
9. SPR	77%	-4.96*	.56** [.31, .76]	Yes	Yes
Total	76%	-18.41*	.81** [.74, .86]	Yes	Yes

Note. CSP = Contemporary School Psychology; IJESP: International Journal of School and Educational Psychology; JASP = Journal of Applied School Psychology; JSP = Journal of School Psychology; PITS = Psychology in the Schools; SPF = School Psychology Forum; SPI = School Psychology International; SPQ = School Psychology Quarterly, SPR = School Psychology Review. Continuous test of the half *p*-curve based on the Stouffer method. As per Simonsohn, Simmons, & Nelson (2015), half curve values that $p < .05$ indicate the absence of right-side bias and thus evidential value. Conversely, evidential value is absent if the power test is $p < .05$ for the half-test.

[a] Outcomes $p < .025$.

* $p < .05$.

** $p > .05$.

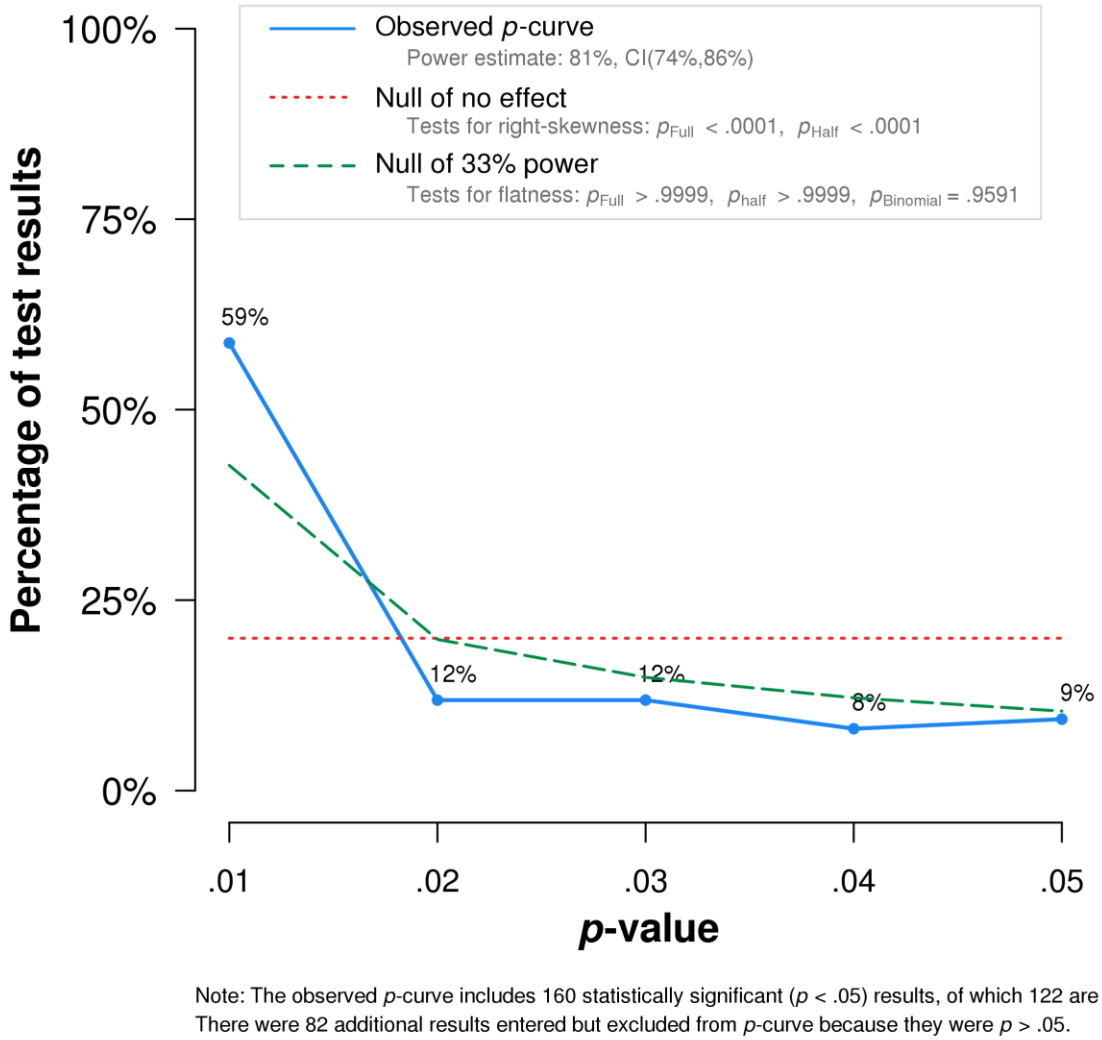

Figure 1. Results from a *p*-curve analysis examining the evidential value of intervention research published in nine school psychology journals from 2016-2017.