aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 537–542

DOI: 10.1177/1745691612460688
http://pps.sagepub.com
SAGE

Replications in Psychology Research: How Often Do They Really Occur?

Matthew C. Makel[1], Jonathan A. Plucker[2], and Boyd Hegarty[3]
[1]Duke University, [2]University of Connecticut, and [3]University of New Hampshire

Abstract

Recent controversies in psychology have spurred conversations about the nature and quality of psychological research. One topic receiving substantial attention is the role of replication in psychological science. Using the complete publication history of the 100 psychology journals with the highest 5-year impact factors, the current article provides an overview of replications in psychological research since 1900. This investigation revealed that roughly 1.6% of all psychology publications used the term *replication* in text. A more thorough analysis of 500 randomly selected articles revealed that only 68% of articles using the term *replication* were actual replications, resulting in an overall replication rate of 1.07%. Contrary to previous findings in other fields, this study found that the majority of replications in psychology journals reported similar findings to their original studies (i.e., they were successful replications). However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles. Moreover, despite numerous systemic biases, the rate at which replications are being published has increased in recent decades.

Keywords

replication, research methodology, content analysis

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

The recent publication of a controversial study on extrasensory perception (Bem, 2011) along with a few well-publicized fraud cases have catalyzed a healthy conversation within the psychological research community about the process used to publish research (e.g., Carpenter, 2012; Crocker & Cooper, 2011; Roediger, 2012; Wagenmakers, Wetzels, Borsboom, & van der Mass, 2011; but cf. Bem, Utts, & Johnson, 2011). One topic that has received substantial attention is the role of replication in psychological science. When do results require replication? Who should conduct these replications? And where should they be published?

Such concerns are certainly not unique to psychology, but they further highlight the importance that replication can play in psychological research. As the introductory quotation by Tukey notes, although replication is far from a miracle cure-all, it can help identify, diagnose, and minimize many concerns about the integrity and reproducibility of research. Some have gone so far as to call replication the Supreme Court or gold standard of science (Collins, 1985; Jasny, Chin, Chong, & Vignieri, 2011, respectively).

However, despite a general positive regard for replications for the improvement of psychological science, conducting replications is viewed as lacking prestige, originality, or excitement (Lyndsay & Ehrenberg, 1993; Neuliep & Crandall, 1993). In other words, a field that replicates its work is rigorous and scientifically sound, but researchers who conduct those replications are looked down on as bricklayers and not advancing knowledge. If the field has truly been set up to deter replications (or to require authors to bend over backward to make their work appear not to be a replication), then one would predict that replications in psychology would be extremely rare. If replications are common, however, their presence would suggest that concerns about disincentives are not warranted. However, to our knowledge, there have been no systematic investigations of the prevalence of various types of replication. The current study sought to inform the discussion of research integrity by investigating replication rates in published psychological research.

Replications in Psychology

The current article provides an overview of replications in psychological research since 1900. We conducted an exploratory investigation into how the issues reviewed above correspond with the publication of replications in psychological

Corresponding Author:
Matthew C. Makel, Talent Identification Program, Duke University, 1121 W. Main Street, Durham, NC 27701
E-mail: mmakel@tip.duke.edu

research. Two primary questions drove our investigation: How many replications are being published? And is the number of replications being published changing over time? We were also interested in whether they were direct or conceptual replication and whether the original findings were successfully replicated.

Lykken (1968) noted that, as researchers, "we are interested in the construct . . . not in the datum" (p. 156). He went on to propose three types of replications: literal, operational, and constructive. Schmidt (2009), in a review connecting the discussion of replication theory with replication practice, eliminated Lykken's (1968) literal replication (because it essentially requires the original investigator to gather data from additional participants) and reframed the latter two types as direct and conceptual replications. In a direct replication, the new research team essentially seeks to duplicate the sampling and experimental procedures of the original research by following the same "experimental recipe" provided in the methods section of the original publication. In a conceptual replication, the original methods are not copied but rather purposefully altered to test the rigor of the underlying hypothesis. Whereas direct replication examines the authenticity of the original data, in conceptual replication, the replicator tests the construct and not the datum to which Lykken referred. We use Schmidt's classification in this article, as it largely encapsulates recent conversations within the field.

Article Selection Process

We identified the top 100 journals according to 5-year impact factor in psychology (all types) by using the online search engine ISI Web of Knowledge Journal Citation Reports, Social Sciences Edition (2010). In May 2012, using Web of Knowledge, we searched the entire publication history of each of these 100 journals to identify the total number of articles published as well as the number of articles that contained the search term "replicat*" in the text (i.e., any articles containing words with the stem "replicat"). This method is similar to what Fanelli (2010, 2011) has used when searching publication histories in large databases.

The replication rate of each journal was calculated (number of articles containing "replicat*" divided by total number of articles) to determine the percentage of articles that were replications. This was also calculated by year, to determine whether the replication rate changed over time. Then, as a reliability check to assess the extent to which the term *replication* was actually referring to a new replication being conducted, we randomly selected and analyzed 500 of the articles containing the term "replicat*." This analysis assessed whether (a) the term was used in the context of a new replication being conducted; if so, (b) whether it was a direct or conceptual replication; and (c) whether the replication was considered a success or failure. The number of times both the replication and original article have been cited was also recorded (if multiple studies were being replicated, the average of the replicated studies was calculated; the citation counts of books were not recorded because they are not calculated by Web of Knowledge). Finally, authorship of each article was also recorded. If the original and replicating papers had no overlap in authors, they were coded as "unique."

All of the data were collected by the first author. The secondary authors were given a set of written instructions (similar to the paragraphs above) to score a subset of articles by using the same procedures as the first author. In 18 out of 20 cases, the articles were assessed similarly, providing evidence that the method identifying replications is itself replicable. The 500 randomly selected articles were then divided and coded independently by the authors.

Analysis

Journal information

The average 5-year impact factor of the top 100 journals in psychology was 5.42 (range = 3.09 to 24.51). Overall, the term "replicat*" was used in 1.57% (5,051 of 321,411) of articles with specific journals ranging from 0% to 6.08% (see the online supplement at http://pps.sagepub.com/supplemental for individual journal data). However, after the year 2000, "replicat*" use was 2.17 times higher (95% CI = 2.06, 2.30) than it was from the 1950s to 1999 (2.39% vs. 1.10%, respectively).[1] This suggests that use of "replicat*" is becoming significantly more common compared with previous decades. In fact, 680 articles used the term "replicat*" in their title (nearly one out of every seven articles that used the term at all). Although this number is quite a small percentage of the overall sample, it is quite high for a field that has supposedly discouraged replications. Using the term in the title is not an indication of trying to hide the concept.

Replication rates

Table 1 reports the results of the more thorough analysis of 500 randomly selected articles that used the term "replicat*." Overall, 68.4% (n = 342) of the analyzed articles that used the term "replicat*" were actual replications.[2] With this correction factor, the replication rate of psychology journals is 1.07%. However, the replication rate did not remain constant. Because of this fluctuation, we added the dashed line to Figure 1, representing the replication rate based on the data analyzed from each decade. Regardless, even after using the correction factors for each time period, the replication rate after the year 2000 was 1.84 times higher (95% CI = 1.72, 1.96) than it was from 1950 to 1999. The increase in replication rate is particularly noteworthy given that it coincides with an explosion in the overall number of articles published. For example, as shown in the secondary y-axis in Figure 1, there were more articles published in the 10-year period from 2000 to 2009 (98,920) than in the entire period from 1900 to 1979 (75,036). Thus, a higher replication rate represents a

Table 1. Replication Rates by Authorship and Success Rates by Replication Type (out of 342 Articles).

	Percentage of replications published		
Replication type	Overall (*N* = 342)	1950s–1999 (*n* = 146)	2000–present (*n* = 196)
In same paper	34.5%	26.0%	40.8%
By same authors	52.6%	47.9%	56.1%
By same journal[a]	19.0%	30.1%	10.7%
Outcome of replications			
All replications		42.7%	57.3%
Successful	78.9%	74.0%	82.7%
Failed	9.6%	15.1%	5.6%
Mixed	11.4%	11.0%	11.7%
Direct (14%, *N* = 48)		13.7%	14.3%
Successful	72.9%	70.0%	75.0%
Failed	14.6%	20.0%	10.7%
Mixed	12.5%	10.0%	14.3%
Conceptual (81.9%, *N* = 280)		80.8%	82.7%
Successful	82.8%	78.0%	86.4%
Failed	7.5%	12.7%	3.7%
Mixed	9.6%	9.3%	9.9%
Both (4.1%, *N* = 14)		5.5%	3.1%
Successful	21.4%	25.0%	16.7%
Failed	35.7%	37.5%	33.3%
Mixed	42.9%	37.5%	50.0%
Replication by authors			
Same authors (52.9%, *N* = 181)		47.9%	56.6%
Successful	91.7%	90.0%	92.8%
Failed	1.7%	2.9%	0.9%
Mixed	6.6%	7.1%	6.3%
Unique authors (47.1%, *N* = 161)		52.1%	43.4%
Successful	64.6%	59.2%	69.4%
Failed	18.6%	26.3%	11.8%
Mixed	16.8%	14.5%	18.8%

[a]This does not count articles that replicated findings within the same article.

multiplicative increase in the overall number of replications being conducted.

Who publishes replications?

Table 1 also reports the percentage of actual replications that were published in the same article (usually through a subsequent experiment reported as part of a multistudy article). Subsequent "follow-up" studies in the same multistudy article may not be as valuable as independent replications (because of potential experimenter bias) but are also not lacking in value. Similarly, an unexpected finding was the high rate of replications being published in the same journal that published the original study; 19% of all replications were published in the same journal as the original study (this does not count 34.5% of replications published in the same article). But perhaps more important, 52.9% of replications were conducted by the same research team as had produced the replicated article (defined as having an overlap of at least one author, including replications from the same publication).

High authorship overlap is important to note because the success rates of replications were significantly different based on whether there was author overlap, with replications from the same research team more likely to be successful than replication attempts from a unique research team (91.7% vs. 64.6%, respectively), $\chi^2(1, N = 303) = 32.72$, $p < .001$, Cramer's V = .33. In fact, when at least one author was on both the original and replicating articles, only three (out of 167) replications failed to replicate any of the initial findings. Such results may reflect the file-drawer problem (i.e., researchers may be loath to publish failed replications of their own work). Although certainly contributing to research knowledge,

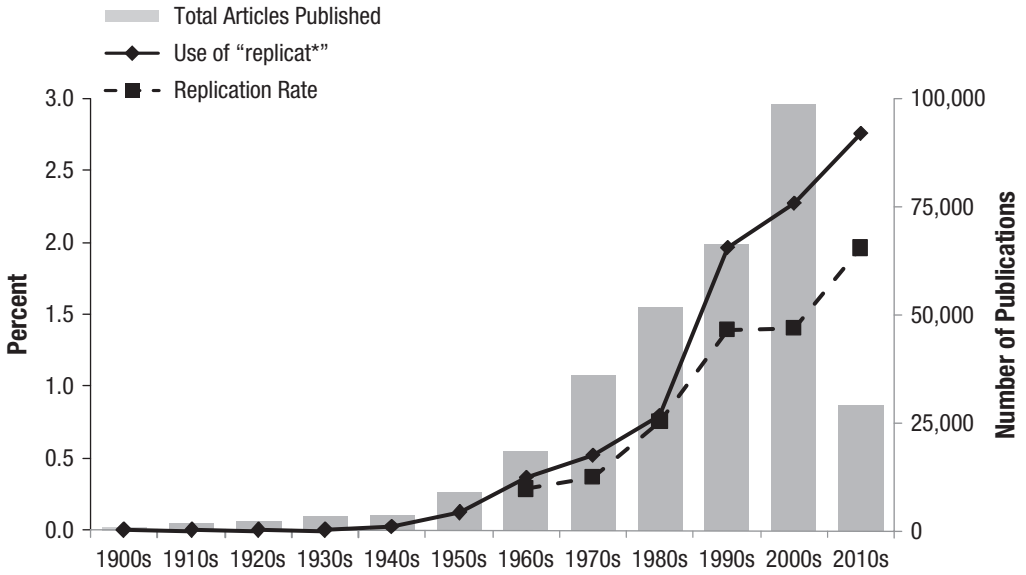

Fig. 1. Replication rate in the top 100 psychology journals. The solid line represents the percentage of publications (from 100 journals with the highest 2010 5-year impact factor) that used the term "replicat*." The dashed line reports the replication rate based on the percentage of articles using the term "replicat*" that were actual replications. The bars represent the total number of articles published in that decade. The 2010s bar is truncated because data from only 2.5 years of the current decade were available.

"first-party replications" may not account for potential experimenter bias, whether intentional or unintentional.

Table 1 also reports the specific breakdown of the success rates of replications, on the basis of whether they were direct, conceptual, or included both (in multiple studies). Overall, 81.9% were conceptual, 14.0% were direct, and 4.1% included facets of both. Although conceptual replications appear more likely to be successful than direct replications (82.8% vs. 72.9%, respectively), this was not statistically significant, $\chi^2(1, N = 295) = 2.94$, $p = .09$, Cramer's V = .10. Although this may seem somewhat counterintuitive (i.e., one would expect successful direct replication to be more likely than conceptual replication), failed conceptual replications may be less likely to be submitted or accepted for publication. Moreover, it may be that only particularly surprising results inspire researchers to attempt to replicate directly.

Citation of replications

The median citation count of the articles that were actually replications was 17 (range = 0–409), whereas the median for the articles being replicated was 64.5 (range = 1–2,099). Obviously, the original articles have had more time to be cited because they are all older than their replicating counterparts are (median publication year of 1992 and 2001, respectively[3]). However, being cited 17 times is quite high (for a comparison, only three of the 100 journals have a 5-year impact factor higher than 17). These citation statistics somewhat weaken the argument that replications are not valued by the research community.

Discussion

The current study sought to provide a comprehensive survey of published replications in psychological research. By analyzing the publication history of the top 100 psychology journals, the current study found that roughly 1.57% of psychology publications used the term "replicat*." A more thorough analysis of 500 randomly selected articles revealed that only 68% of the articles that used the term were actually replications, creating an overall replication rate of 1.07%. Contrary to previous findings in other fields (e.g., Ioannidis, 2005), this study found that the vast majority of both direct and conceptual replications in psychology journals reported similar findings to their original studies (i.e., successful replications). However, replications were significantly less likely to be successful when there was no overlap in authorship between the original and replicating articles.

As seen in Figure 1, an inflection point appears in the current data in the 1990s, with a significant jump in replication rate. It is interesting to observe that the growth in replications over time (i.e., the slope of the dashed line in Fig. 1) flattened between the 1990s and 2000s but appears to be increasing in the 2010s. This may be a function of the recent increased attention to positive bias, the file-drawer problem, and prevention of scientific fraud. The replication rate found in the current study is not dissimilar to replication rates reported in other fields. Although comprehensive replication rates could not be found for other domains, individual studies report that replication rates in business, marketing, and communication journals range from 1% to 3% (Evanschitzky, Baumgarth, Hubbard, &

Armstrong, 2007; Hubbard & Armstrong, 1994; Kelly, Chase, & Tucker, 1979). However, unlike the current study, those investigations reported apparent slowdowns in replication rates over time.

What merits replication?

Currently, the system used to determine what studies merit replication is, as Hunt (1975) somewhat generously described, "informal and somewhat haphazard and could be improved" (p. 588). On an intellectual level, it is a question of optimizing the relationship between resources devoted to research and the accuracy of results. On a practical level, replicating important and relevant findings provides policy makers with important information needed to create effective policy. However, at the same time, there is concern over who is responsible for determining what needs replicating (and who should do it). As shown in the current article, the original research article need not even have spurred hundreds of citations for a replication to get published. Nevertheless, as a field, we need to decide what is "good enough" for replication rates. We are reluctant to provide a recommendation for how many replications should be published because other recommendations, such as using *p* values less than .05 and Cohen's demarcation of small, medium, and large effect sizes, are typically misused and interpreted more like laws rather than the cautious suggestions they represent (e.g., Cohen, 1988; Rosnow & Rosenthal, 1989). We are not suggesting that every undergraduate thesis requires replication, but a conversation about the replication of important studies that impact theory, important policies, and/or large groups of people would provide useful and provocative insights, particularly via the implementation of modern methods and measures. That being said, as an arbitrary selection, if a publication is cited 100 times, we think it would be strange if no attempt at replication had been conducted and published. Such a guideline would help avoid flawed or fraudulent findings going unquestioned over an extended period of time. Research findings require replication because of their influence, not despite it.

Caveats

If research articles are not framed as replications, then they were not categorized as such. A potential limitation to the current study is that if an author actually intended to replicate but did not explicitly include that term, it was not captured by the methods used in the current study (cf. Kelly et al., 1979). Of course, this limitation extends beyond the present study; it also limits readers in their ability to connect research to its intellectual precedents. To calculate the rate of replications relying on a cloaking device, the entire library of articles would have to be analyzed, which is impractical. Similar in-depth investigations would be required to calculate accurate replication rates of specific journals. Future research delving more deeply through entire issues and volumes of journals may reveal higher replication rates (not to mention different success rates) once hidden replications are unmasked, but we doubt the conclusions of that multiyear study would be much different from those emanating from the present methods.

Finally, it is important to remember that replications are not a cure-all. Just as Campbell and Stanley (1963) cautioned against considering experimental methods as a panacea, all replications are not of equal value. In particular, conceptual replications published within multistudy articles do not necessarily satisfy all the goals of replication, including limiting experimenter or measurement bias. And, of course, failure to replicate does not necessarily suggest that a research team is in the wrong.

The proverb "No press is bad press" is certainly difficult to swallow when the field is receiving so much negative attention and criticism about the quality and integrity of its research. But if the field comes out of the current scrutiny with stronger, more rigorous methods that lead to deeper understanding of psychological constructs, then this scrutiny is a much needed and welcomed push forward. Whether it be in initiation of incentives or the removal of roadblocks, the path to better understanding psychological science goes through replicating important research findings.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. The first use of the term *replication* that we found in a psychology journal was Rosenblith's article (1949) in the *Journal of Abnormal Psychology*, titled "A Replication of Some Roots of Prejudice," which successfully replicated the findings of Allport and Kramer (1946) while relying on college students in South Dakota instead of those in Harvard, Radcliffe, and Dartmouth.
2. The articles that used the term "replicat*" but were not actual replications typically used related terms in the context of stating that the study's results needed to be replicated, that more replications of a given study were needed, or that specific genes were replicated, or a specific database with "Replication" in its name was used in the given study (e.g., National Comorbidity Survey Replication).
3. Only articles that replicated previously published findings were included in this comparison; articles that replicated only another study from the same article (i.e., a multistudy article) were excluded.

References

Allport, G. W., & Kramer, B. M. (1946). Some roots of prejudice. *Journal of Psychology*, *22*, 9–39.

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

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716–719. doi:10.1037/a0024777

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, H. M. (1985). *Changing order*. London, England: SAGE.

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

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. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. doi:10.1007/s11192-011-0494-7

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

Hunt, K. (1975). Do we really need more replications? *Psychological Reports*, *36*, 587–593.

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

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

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

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

Lyndsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician*, *47*, 217–228. doi:10.2307/2684982

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

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

Rosenblith, J. F. (1949). A replication of some roots of prejudice. *Journal of Abnormal and Social Psychology*, *44*, 470–489.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284. doi:10.1037//0003-066x.44.10.1276

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

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. *American Psychologist*, *24*, 83–91. doi:10.1037/h0027108

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790