Major Factors Affecting the Non-Replicability Problem in Psychology and Other Sciences

By

Lewis D. Eigen

Created: November 4, 2016

Last Edited: May 14, 2017

**For Review Only Please
Comments & Suggestions Are Welcome**

Contact

E-Mail: leigen@eafoundation.com

Phone: (301) 983-022

Abstract

This article provides a partial explanation of the non-replicability problem that has recently embarrassed the field of psychology and other research fields and caused many to lose confidence in our research methods and findings. Specifically it demonstrates that there are other factors contributing to the problem other than the “bad apples” of dishonest or inadvertently biased researchers. This paper demonstrates mathematically that a very significant part of the non-replicability problem is statistical in nature and is inevitable as long as the existing systems and norms of academic, scientific publishing are as they are. It is demonstrated that the very selectivity of the editorial boards of the refereed journals dramatically effects the probabilities that almost all statistically based hypothesis testing research is based. Another factor is the testing of multiple hypotheses in a single study which also induces error for purposes of replicability into our standard statistical tools and techniques. This paper strongly suggests that the very refereed journal system we have in place to ensure the quality of research in the field, in actuality, does just the opposite, and undermines it and then adds its prestige and imprimatur to the validity of the study despite its inadvertent eroding the quality and the confidence that should be placed in the study.

Introduction

What is the reason for the non-replicability of so many studies in psychology and other social sciences? To be sure we have our share of “bad apples” who falsify results and other dynamics—such as commercial sponsorship of research--that might tend to bias investigators’ judgment. However, this analysis shows that these may be only a minor factor. More responsible for the undesirable state of non-replicability are factors inherent in the very structure of how we conduct and disseminate research findings. Further, I submit that the same problem exists in all the social sciences, education, medicine, and other fields which use statistical methods for hypothesis testing and have similar professional publication norms. If replications in those fields, as well as psychology, were attempted on a large scale, a substantial number of the peer-reviewed, published studies would also fail replication. However, for various well-known reasons relatively few published studies are replications. In a study of the leading 100 journals of psychology,

only 1.6 percent of all publications were replications¹. A similar finding occurred in the education field.²

Other than “bad apples” and bias, there are two other major factors that have caused this problem and will continue to cause the problem unless we make some changes in the way we conduct, report and disseminate our research. The research referred to in this analysis does not include case studies or other types of research that do not include statistical testing of hypotheses.

The First Factor: Research Publishing Norms

Our academic, professional publication system was originally designed to improve the quality of published studies and assure the various publics that published studies have met the highest standards of contemporary methodology and best practices and have been subjected to peer review. However, there has been a major unintended and unexpected consequence: IT actually is responsible for a substantial proportion of the problem of non-replicability.

Essentially, most of our statistical research involves the testing of a null hypothesis and rejecting same if and only if whatever differences are found are “statistically significant”. Typically, we use the $p=.05$ level of significance as our threshold for rejection, but occasionally we see $p=.01$ and $p=.10$ used also, as well as other levels in between. For the rest of this article. I will assume that the .05 level of significance is typically used, but the mathematics supporting my thesis demonstrates the points regardless of the level of significance that is used.

If a graduate student would gather 15 different elements of data on a set of subjects, and then calculate a correlation matrix of the data, he/she would end up with 210 different correlation coefficients. And if 10 of these were statistically significant, the student may become excited over his significant findings, but we would use that situation to teach the student the pitfalls of doing many analyses—statistically testing many hypotheses and picking out those that were significant to draw conclusions. We would explain that if one obtains 10

¹ Makel, Matthew C.; Plucker, Jonathan A.; Hegarty, Boyd (2012-11-01). "[Replications in Psychology Research How Often Do They Really Occur?](#)". *Perspectives on Psychological Science*. 7 (6): 537–542. doi:10.1177/1745691612460688. ISSN 1745-6916. PMID 26168110.

² Matthew C. Makel and Jonathan A. Plucker, “Facts Are More Important Than Novelty: Replication in the Education Sciences”, *Educational Researcher*, August 14, 2014

correlations that are statistically significant at the .05 level, that is just what we would expect very often if there were no real relationships among the variables.

Likewise, if 200 researchers conducted one study each, producing 200 different studies, each testing a single null hypothesis and **none** of those hypotheses really had any underlying difference, we would still expect slightly more than 10 null hypotheses to be rejected—sometimes more and sometimes less -- to be “significant” at the .05 level. But we would, if we repeated this many times, expect an average of exactly 10.1 of the studies to show significant differences when in point of fact there were no underlying differences. We will have a number of null hypotheses spuriously rejected. It follows that if some other researchers focused on the 10 or so “significant” findings and tried to replicate them, the vast majority of the time the replication would fail. On the average, only about $\frac{1}{2}$ of one single study would be “replicated”—as the probability of having two spurious findings with the same study would be .0025 assuming that the .05 significance was used for both.

In general, many experimental psychologists and others using modern statistical procedures, recognize this danger, and although we are enthused by significant findings in any single study, most of us realize that there is roughly a 5 percent chance that the finding at the .05 level was spurious. That is one of the reasons that we like to see replications which will enable us to have much more confidence that we are rejecting a null hypothesis because of the dynamics of the data and not the dynamics of the statistical analysis of the data. And this illustrates why replication of studies is so important in any science that utilizes modern statistics. The chance of obtaining a spurious rejection of a null hypothesis is 5 percent, while if all studies were replicated, the odds of having a spurious finding would only be $\frac{1}{4}$ of 1 percent.

If a scientific journal publishes a set of studies where each rejects a null hypothesis (at the .05 level of significance), we should statistically expect that 4.875 percent of the studies published in that journal would fail to replicate (5 percent less the small percentage that would be spurious twice).

So if in the recent replication attempts only 5 percent or so of the studies failed to replicate, we would not be alarmed. That is to be expected. However, the replication failure percentage has been, and is, **much higher**—32 percent and higher depending on definitions of replicability.^{3,4} “Bad apple” researchers and researcher bias and methodologic error⁵ will add to the 5 percent proportion, but the field of psychology would have to have a very high percentage of methodological stumblebums or unethical and biased investigators to account for the huge difference. But that sad state of affairs is certainly a logical possibility. Today, too many people in our society believe this explanation, and some politicians use this explanation to reject the findings of science in favor of their ideology of the day. However, *Occam’s Razor*, a sound guide in all research and analysis, tells us to look for the simpler solutions first, and until they are rejected, do not consider and test the more complex explanations. There are, however, several possible simple explanations of the very high, non-replicability frequency and so these are examined herein.

It is recognized by almost all experimental researchers that there is a great deal of bias in the determination of what does and what does not get published in our peer reviewed journals. I do not refer to cronyism, institutional prejudice and the other forms of bias about which most editorial boards and editors are constantly vigilant. I refer to what some call the “bias of positive results” which I prefer to label “the tendency toward positive results”. The reason “tendency” is a better term, is that “bias” has a negative connotation for most. The tendency toward positive results is arguably very desirable and important in setting research agendas. The late Columbia University experimental psychology professor, Nat Schoenfeld, when I was a student in one of his advanced research courses, castigated one of my fellow students for suggesting that a study which did not have positive findings was none-the-less, quite valuable because it provides us an opportunity to learn what was not the case. Shoenfield said,

Well trained experimental psychologists are too precious and valuable to waste their time looking for crumbs of knowledge in such a study. You certainly cannot conclude that the null hypothesis was true. And there are always myriad reasons that the experiment, as conducted, might not have shown positive results even if they were there. The instruments and

³ Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012)

⁴ Open Science Collaboration. *Science* <http://dx.doi.org/10.1126/science.aac4716> (2015).

⁵ With the advent of PCs running statistical packages at little effort or cost, a common methodological error that is being made all too often is that the researcher runs a panoply of different statistical tests and if one turns out to be “significant” that is the one that is selected for the write up of the study. Often this use of multiple statistical methods and tests is not reported in the publication that ensues.

measurements that were used might not have been sensitive enough, the experimental procedures could have been faulty, the sampling likewise, the calculations could have been erroneous—there are dozens if not hundreds of possibilities, and we cannot have highly trained experimental psychologists wasting their time and society's limited research funds looking through that pile of horseshit in the hopes that they will find a pony. The only ones who should be examining the non-positive results study, are the investigating team and their faculty advisors or supervisors. The rest of us have much more important things to do with our time!

He went on to explain that is why we have peer-reviewed journals—to not only assure sound research procedures were used, but to filter the positive result studies from which we can learn.

However, despite Shoenfield's argument which has, to a great extent, characterized that of almost all the journal editors and editorial boards in creating the tendency toward positive results, there appear to have been a serious unintended consequence of this tendency that bears on the non-replicability problem. This can be demonstrated with a thought experiment.

Given that almost all researchers are aware of the tendency towards positive results, most investigators who conduct studies and do not show positive results simply do not submit them for publication. If there are a few that do submit, the chances are that they get rejected. So in our thought experiment, let us consider all studies with positive results that are submitted and are accepted for publication⁶.

Assume that Ω studies are embarked upon and after completion. Ω_1 show positive results (significant at the .05 level) and Ω_0 do not where $|\Omega_0| + |\Omega_1| = |\Omega|$. To determine the percentage of studies with spurious results compared to all those that are that get published, we divide 5 percent of all the studies that were conducted ($.05 \times |\Omega|$) by the total number of published studies-- Ω_1 .

⁶ Clearly this is not so, but it just makes the example a little clearer.

If roughly half of all studies have positive results and are published, then the percentage of spurious studies of **all published studies** will be almost 10 percent and not 5 percent. This means that nearly 10 percent of all the studies will likely not replicate instead of the 5 percent that we intuitively believe the percentage will be.⁷

In this case we assumed that the ratio of published studies with positive results to all studies conducted was .5 (½). If it were smaller, then the percentage of spurious studies to all would be higher still. And of course, if the proportion of all studies that were published to all studies conducted were larger, then the percentage would be less, but it is always more than the intuitive 5 percent that would go with the .05 significance level.

Table 1
Average Percent of Spurious Significant Studies

		Average Significance Level									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Publication Ratio – Published Studies	0.1	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
	0.2	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
	0.3	3%	7%	10%	13%	17%	20%	23%	27%	30%	33%
	0.4	3%	5%	8%	10%	13%	15%	18%	20%	23%	25%
	0.5	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
	0.6	2%	3%	5%	7%	8%	10%	12%	13%	15%	17%
	0.7	1%	3%	4%	6%	7%	9%	10%	11%	13%	14%
	0.8	1%	3%	4%	5%	6%	8%	9%	10%	11%	13%
	0.9	1%	2%	3%	4%	6%	7%	8%	9%	10%	11%
	1.0	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%

⁷ The reason that it is not exactly 10 percent is that there is a small probability that a spurious, significant finding, when replicated, will again produce a spurious result.

Table 1 provides the real percentage of spurious studies that will not replicate, based on two variables. The horizontal axis (columns) which shows the average significance level that is required of the studies to reject a null hypothesis and the vertical axis (rows) is the ratio of published studies to all conducted – the Publication Ratio. This is not a readily available statistic. It is much smaller than the average rejection rate of the journals, as many studies are never submitted for publication. We can only speculate for now⁸, but the findings of this analysis are remarkably consistent regardless of the magnitude of this statistic.

The former scale ranges from .01 to .1 in .01 increments while the latter scale runs from .1 to 1.0. The 1.0 would be the case if ALL studies were published in refereed journals, and ironically this practice would eliminate this particular problem that I have identified.

At present, it is impossible to tell what a realistic estimate should be for selecting the Publication Ratio. While different Journals have different percentages of acceptance, that statistic has little or nothing to do with the publication ratio. However, a very good estimate of the current publication ratio could be obtained by conducting studies for that purpose, which we submit is a very important study to be done for the health and credibility of our entire profession.

We have the ironic and counter-intuitive situation whereby the very efforts our research journals have made to increase the quality and validity of the research that is promulgated to the field has actually, and unintentionally, produced another kind of error – namely the publication of many studies that will not be able to be replicated.

This is also a classic example of the paradox of the commons: When all can use a common facility such as a pasture or a waste dump, if the facility is good and desirable, then there will be more use and some abusive use. The costs of the commons is borne by the community, but individuals glean the benefits. So, if I use more than my fair share of the commons for grazing my cattle, then I benefit proportionately more. And then the rest of the community must not only pay

⁸ Prior to this kind of analysis, there was not a great deal of utility to knowing this ratio, but it now turns out to be a very material factor in the replicability problem. It is my opinion, a survey of researchers at all levels (including undergraduates and amateurs) would be of significant value in further analysis of the non-replicability problem.

more than their share of the costs, but may have to ration everyone else's grazing privileges to accommodate mine. In a commons, that which is the best strategy for the individual users may be the opposite for the community as a whole. In the instant case, each journal produces higher and higher standards of publication to make their particular journal more credible and respectable, and the result of all the journals doing this is a lack of credibility and respectability for the field as a whole. In the long run, the journals suffer perhaps the most—it is **their** vetted studies that will not replicate. And the higher their standards, the more likely that their published studies and their editorial staff and board will be tarred by non-replicability.

Fortunately, there are methods of managing the commons that have been developed to optimize benefits and minimize unintended adverse consequences, and later in this article I will describe some possibilities.

The Second Factor: Multiple Hypotheses

In physics and chemistry research, almost all studies deal with a single hypothesis. A research program may have a number of studies, but each typically has only a single hypothesis. Not so in experimental psychology or most of the social and medical sciences. There is no current source of data known to me to determine the average number of hypotheses that are tested per study in experimental psychology or other research fields.⁹ One could sample Journals to get a good estimate, but as should be clear by the preceding arguments, the estimate to be useful must be made for all studies, published and not. However reasonable informal discussions with some colleagues in the field for a number of years—especially those who have supervised numbers of doctoral dissertations would tell us that the number is certainly more than 1 and might be considerably higher. In 1962 Cohen¹⁰ reported surveying 70 studies that year and calculated an average of over 70 different statistical tests per study. Alex Reinhardt¹¹, who has been studying the multiple hypothesis testing problem for a number of years, is of the opinion that the average number of hypotheses tested is “very much larger” than the 2 that I have used for purposes of this argument. One advantage to assuming only 2 or 3 as I do herein is that these is

⁹ This is another survey area that would improve the understanding of non-replicability.

¹⁰ Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.

¹¹ Personal communication, July 2017

certainly a lower bound in the real average number of hypotheses tested. At this time, just assuming that it might be 2 or 3 will give us insight into this second factor that also accounts significantly for the non-replicability problem.

If a study has two null hypotheses to test, how would the principle investigator determine if he/she had obtained positive results? Most of us would declare positive results if we could reject at least one of the 2 null hypotheses. And many journals have published such studies where one or both null hypotheses have been rejected. Both rejected is even better, but studies rarely get published if no hypotheses are rejected. Likewise, if there were 3 hypotheses, a positive result might be claimed if only one of the three were rejected at the specified level of significance. 2 or 3 would be even better.

Now consider the effect on the number and proportion of spurious positive results. If there were 2 hypotheses to be tested, then, if only 1 rejected hypothesis is needed for a positive finding study, then the proportion of such studies that would result in a spurious positive result would not be .05 if .05 were the level of significance required. It would be .975 or almost double--ten percent. With 3 hypotheses, it would be just under 15 percent. These would not usually replicate.

If we then consider the effect of the two factors together—the Publication Policy factor and the Multiple Hypotheses factor, then **20 to 30 percent of all the published studies would have a spurious positive finding and would not likely replicate.** This seems to be a close approximation to the initial rough estimates of non-replicability that have recently been reported. It is a few percentage points lower, and any realistic analysis of this non-replicability problem must allow a few percentage points for skulduggery and incompetence.

Table 2 illustrates the average percentage of spurious significant studies if there were two hypotheses tested in the same study. And Table 3, for 3 hypotheses.

Table 2
Average Percent of Spurious Significant Studies
2 Hypotheses Tested

		Average Significance Level									
Publication Ratio – Published Studies to All		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
	0.1	20 %	40 %	60 %	80 %	100 %	120 %	140 %	160 %	180 %	200 %
	0.2	10 %	20 %	30 %	40 %	50%	60%	70%	80%	90%	100 %
	0.3	7%	13 %	20 %	27 %	33%	40%	47%	53%	60%	67%
	0.4	5%	10 %	15 %	20 %	25%	30%	35%	40%	45%	50%
	0.5	4%	8%	12 %	16 %	20%	24%	28%	32%	36%	40%
	0.6	3%	7%	10 %	13 %	17%	20%	23%	27%	30%	33%
	0.7	3%	6%	9%	11 %	14%	17%	20%	23%	26%	29%
	0.8	3%	5%	8%	10 %	13%	15%	18%	20%	23%	25%
	0.9	2%	4%	7%	9%	11%	13%	16%	18%	20%	22%
	1.0	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%

When there are 3 hypotheses, the effect is even stronger and there will an even larger proportion of spurious studies. Since 2 and 3 are lower bounds for the real average number of hypotheses tested makes the situation even worse.

Following is Table 3, the results if there were 3 hypotheses being tested.

Table 3
Average Percent of Spurious Significant Studies
3 Hypotheses Tested

		Average Significance Level									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Publication Ratio – Published Studies to ...	0.1	30%	60%	90%	--	--	--	--	--	--	--
	0.2	15%	30%	45%	60%	75%	90%	--	--	--	--
	0.3	10%	20%	30%	40%	50%	60%	70%	80%	90%	--
	0.4	8%	15%	23%	30%	38%	45%	53%	60%	68%	75%
	0.5	6%	12%	18%	24%	30%	36%	42%	48%	54%	60%
	0.6	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
	0.7	4%	9%	13%	17%	21%	26%	30%	34%	39%	43%
	0.8	4%	8%	11%	15%	19%	23%	26%	30%	34%	38%
	0.9	3%	7%	10%	13%	17%	20%	23%	27%	30%	33%
	1.0	3%	6%	9%	12%	15%	18%	21%	24%	27%	30%

This analysis suggests that there are at least three factors producing the non-replicability problem: Factors 1 and 2 from this paper, and the “bad apples” and biased researchers. If the negative researcher characteristics are as high as 5 percent (and I think that is a very high estimate from my personal experience)--but if it were, we could easily be looking at **one third of all published studies today being unable to be replicated.**

Note that even under these assumptions, dishonesty, bias, corruption and incompetence would represent less than one-sixth of the non-replicability problem. Five-sixths would be these statistical inevitabilities brought about by our journal publication norms and multiple hypothesis testing.

Many professional fields try and cover up and minimize the dishonesty and corruption that exists among their professional members. If this thesis herein is correct, we have **amplified and exaggerated our dishonesty and corruption by a factor of six.**

What Might Be Done To Ameliorate

Without much discussion, I offer the following set of suggestions for our field. I do not suggest such “Draconian” methods of refereed journals giving up or lowering their standards for publication. Although it might be a good idea for the commons, human nature and our professional training and values would make that very unlikely to happen. Nor do I make suggestions that would require a major change in the way we conduct our research. These are small changes that could produce big results.

1. **Experimental Design History and Details:** Existing and new journals could add a requirement for articles which, for lack of a better term, I will call “Experimental Design History and Details”. This should include a history of the particulars of the experimental design such as the number of hypotheses tested, when it was decided to test those hypotheses (after the data had been inspected or before). Was the design based in a pilot study and if so give a brief description? As one example of a problem that would be ameliorated, consider the researcher who tests 3 hypotheses, and finds that 2 of them produce significant results. He/she might then write up the study only discussing the two rejected hypotheses and leaving out any mention of the third. This type of description of the design history of the study is standard in many physics and chemistry journals.
2. **Adjust Significance Tests to Account for Multiple Hypotheses:** Establish a norm whereby researchers adjust the mathematical formulas they use to test hypotheses to account for the necessary correction when there are multiple hypotheses tested in a single study. It would be relatively easy for

the computerized statistical packages to add such adjustments in future versions. Initially a publication issued by AAAS for some other prestigious organization can simply produce a set of tables regarding the adjustments that should be made for multiple hypotheses testing. For more mathematically sophisticated researchers, there are well known ways of making these corrections. A few currently use them, but the vast majority do not at this time.

3. **Establish a Completed Study Data Base (CSDB) for Each Field of Research:** Create a website and a national (perhaps international) data base of all studies done in the field¹²--a Completed Study Data Base (CSDB), to include all studies published and unpublished. The website can be designed so that the principle investigator can make the entry on-line herself. The record of a study requires minimal information: Title of Study, Names and contacts of Investigators, Date of Completion, references to sources of published results if any, and a brief description of the study—the abstract. In addition, there would be a binary field to indicate whether or there were statistically significant findings in the study.

Such a web data base could go one step further and allow each researcher to post a write up on his study even if it is not submitted for publication or has been rejected. There is an excellent exemplar of an on-line data base of unpublished studies operated by The Center for Open Science—OSF Preprints¹³--which only lacks the basic form for the author to enter some critical data so that the statistics for each field can be aggregated and reported in real time

A simple computer program on the web server of such a data base can produce a dashboard page on the website to report (for any period selected by the user), the total number of studies performed, the proportion where there was a statistically significant positive outcome, and the total number of studies and proportion where there had been a publication in a refereed journal—the Publication Ratio. In addition, there would be a field indicating whether or not the study was a replication and whether or not the replication was successful. Given those data, the computer can calculate and display

¹² There would best be separate data bases for psychology, political science, economics, sociology, health & medicine, etc. However, within each, data can be separately for the different sub-fields within larger fields such as psychology or Medicine.

¹³ <https://osf.io/preprints/> The major current function of this website is to allow researchers to circulate drafts of papers of ideas for studies to obtain critical feedback and suggestions from colleagues. The site serves most of the scientific fields and has over 190,000 users registered in psychology alone.

the reasonable expected proportion of spurious positive findings for that period of time for that field.

With that information being publicly available, professionals in the field will have a much better sense of the likelihood that a study done in that period could be replicated. Further, over time we will know the percentage of attempted replications that were successful as opposed to those attempts which failed to replicate. By comparing this proportion to the expected one above, the difference will be a very good estimate of how much researcher dishonesty or bias may be occurring. And that will be visible on the dashboard and transparent to all. By following these statistics over time, we can estimate if the investigator dishonesty/bias proportion is decreasing or not. We, as well could estimate the order of magnitude of how serious that problem might be. Further, for the future, when a researcher reports a failed attempt at a replication, that will not be so alarming to the professional and lay public as has been the case so far.

Of the three suggestions herein, this last is the only one that requires substantial resources. All 3 require professional will to implement, but the National (International) data base is the one that will require much more than will. To accomplish number 3, I would propose that the data base be established and maintained under the auspices of AAAS—that a committee be appointed to specifically define the requirements for the generic **Competed Study Data Base (CSDB)**. The requirements should include the data to be collected and the specifics of the Dashboard that will make the whole process transparent. The programming of the website and the data base will require significant resources—perhaps \$500,000¹⁴. Serving the website will also have an annual cost—perhaps \$50,000¹¹. One real possibility is to approach some of the large digital companies such as Microsoft or Google and ask them to program and serve the data base as part of their pro bono work.¹⁵ Another is to “piggy-back” on the existing Center for Open Science website described above.

For the many Principle Investigators conducting studies, the burden to enter their studies in this data base would not be substantial. Perhaps 10-15 minutes of

¹⁴ This is based on the author's personal experience with creating and deploying similarly complex data bases for public web access.

¹⁵ Google, in particular, would find this project an excellent fit for its mission and existing activities. They have an existing policy whereby professional employees are allowed one day a week to work on projects in which they might be interested. Many would likely be interested in such a project. Further, if the AAAS committee were to go a little further and allow PIs who enter their studies in the data base, to also attach reports and or data sets for dissemination to the field.

time. It would help if professional organizations like the American Psychological Association and the AAAS were to pass resolutions encouraging entering all studies in the **CSDB** for each field of science where statistically based studies are common. Further, it would be very desirable if the peer reviewed journals would include requirements that published studies be also registered in the data base. Granting and funding agencies can also make **CSDB** data base entry a requirement of funded research. And universities could make such entry of the study a requirement for a masters of doctoral dissertation as well as undergraduate research when it takes place.

Such a **CSDB** for each scientific field could serve as another research tool for the different research fields, if there were some classification of tagging of each study entry so that they could be classified and analyzed with finer granularity. For example, there may be very different ratios with learning studies as opposed to perception studies within the field of psychology. But this is also something that should be in the purview of the AAAS committee.

There are doubtless other steps that could be taken to ameliorate the non-replicability problem as we are just beginning to understand the factors that cause the problem.

Authors' Note

The analysis herein has very profound implications for the social sciences, their research, reputation, research dissemination modes, and the utility of social and medical sciences to our society. These results and their implications were, candidly very surprising to me, and without my original training as a mathematician, these relationships would probably have never occurred to me. They also imply great change for our profession.

From my original insights on this issue in November of 2016 to this version of this preprint in August of 2017, I have sought out as many reviews and suggestions from colleagues, always considering the possibility that I may have missed something very critical in my analysis. I was candidly fearful that if I published this and there was a major error of logic or mathematics, it would appear the non-scientific part of our population as if this was a “whitewash” of a corrupt, incompetent, and not really scientific profession.

I therefore welcome any suggestions, criticisms, improvements and or additional ideas for amelioration. Feel free to circulate the contents herein as well as any thoughts and materials that may be suggested by it. And feel free to reference it or even quote from it, put please cite the fact that this is a preprint and a working draft.

I thank all in advance for any assistance to me and our profession that you might render.