ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(3) 333–342
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614529796
pps.sagepub.com
SAGE

# Continuously Cumulating Meta-Analysis and Replicability

**Sanford L. Braver[1], Felix J. Thoemmes[2], and Robert Rosenthal[3]**
[1]Arizona State University & University of California, Riverside; [2]Cornell University; and
[3]University of California, Riverside

**Abstract**

The current crisis in scientific psychology about whether our findings are irreproducible was presaged years ago by Tversky and Kahneman (1971), who noted that even sophisticated researchers believe in the fallacious Law of Small Numbers—erroneous intuitions about how imprecisely sample data reflect population phenomena. Combined with the low power of most current work, this often leads to the use of misleading criteria about whether an effect has replicated. Rosenthal (1990) suggested more appropriate criteria, here labeled the *continuously cumulating meta-analytic* (CCMA) *approach*. For example, a CCMA analysis on a replication attempt that does not reach significance might nonetheless provide more, not less, evidence that the effect is real. Alternatively, measures of heterogeneity might show that two studies that differ in whether they are significant might have only trivially different effect sizes. We present a nontechnical introduction to the CCMA framework (referencing relevant software), and then explain how it can be used to address aspects of replicability or more generally to assess quantitative evidence from numerous studies. We then present some examples and simulation results using the CCMA approach that show how the combination of evidence can yield improved results over the consideration of single studies.

**Keywords**

replication, meta-analysis, statistical intuition, effect-size heterogeneity

Over 40 years ago, future Nobel Prize winner Daniel Kahneman and Amos Tversky[1] proved their perspicacity by anticipating the current crisis of replicability—the doubt that the same or another researcher can reproduce (repeatedly) the same phenomenon or empirical relationship with subsequent empirical investigations, implying that the phenomenon may be illusory. Presciently, in their 1971 *Psychological Bulletin* article entitled "Belief in the Law of Small Numbers," Tversky and Kahneman wrote that "most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding" (p. 105). Their article title was a play on The Law of Large Numbers. This well-proved law dates back to Jacob Bernoulli in 1713, and "guarantees that very large samples will indeed be highly representative of the population from which they are drawn" (Bernoulli, 1713, p. 106). But Tversky and Kahneman noted that humans, even highly sophisticated researchers, have "strong" but "wrong" intuitions that "the law of large numbers applies to small numbers as well" (p. 106).

In the last 2 years, these powerful but misguided intuitions have contributed to a profound reconsideration of social science practice and culture. The replicability of the totality of psychology's findings is now being questioned and challenged. Critics note that disturbingly few successful replication studies are in our literature (Carpenter, 2012; Makel, Plucker, & Hegarty, 2012; Open Science Collaboration, 2012; Yong, 2012; but see the "many labs" project, Klein et al., 2014). This serves as proof either that our phenomena are fundamentally false (i.e., those replication studies that have been attempted fail to produce supportive findings) or that journals are biased against replication studies whatever their result (Giner-Sorolla, 2012; Neuliep & Crandall, 1990, 1993; Rosenthal, 1979). This feeling of mistrust was further

**Corresponding Author:**
Sanford Braver, Arizona State University, Department of Psychology, PO Box 876005, Tempe, AZ 85287-6005
E-mail: sanford.braver@asu.edu

fueled by high-profile failures to replicate effects (see e.g., Bargh, Chen, & Burrow, 1996, and the subsequent failed replications by Doyen, Klein, Pichon, & Cleeremans, 2012; Harris, Coburn, Rohrer, & Pashler, 2013). The ultimate result, some conclude, is that the "scientific literature is too good to be true" (Bakker, van Dijk, & Wicherts, 2012, p. 543; Francis, 2012); one cannot believe anything published, and psychology is filled with mostly invalid and false explanations. An article in *The Chronicle of Higher Education* (Bartlett, 2012) was titled "Is Psychology About to Become Undone?"

In this article, we argue that the erroneous belief in the Law of Small Numbers, in combination with the typically low levels of statistical power of psychological (and other similar types of) research, has contributed to this crisis by leading to inappropriate criteria applied to deem an attempt at a replication unsuccessful. We present an alternative and, we believe, superior way to assess evidence of replication.

## The Criterion of Successful Replication

Tversky and Kahneman (1971) posed the following hypothetical question to sophisticated researchers, including members of the elite Mathematical Psychology Group:

> Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group? (p. 105)

> The median answer they obtained from their panel of expert researchers was .85, which was almost double the actual probability (which they show to be about .48). The authors attribute their respondents' extraordinary inaccuracy to the ingrained but fallacious idea that a rather small sample "randomly drawn from a population is highly representative, that is, similar to the population in all essential characteristics." (p. 105)

The correct general answer to such a hypothetical is precisely whatever is the statistical power of the test performed in the replication study. Power is commonly designated 1-β (where β is the probability of a Type II, or false negative, error). Power is generally determined by four factors: (a) the population effect size, which is technically unknown, but is commonly estimated or guessed at; (b) the total sample size, *N*; (c) the "sidedness" of the test (whether one tailed, directional; or two tailed, nondirectional); and (d) the chosen alpha level, which in psychology is typically fixed at .05.[2] Cohen (1962) estimated that the typical power of published psychological studies does not exceed 0.5; Sedlmeier and Gigerenzer (1989) found a similar value among studies conducted in the subsequent 25 years, indicating that the low power of most earlier published behavioral research was not rectified by Cohen's revelation. It has apparently escaped recognition by most that, with the low levels of power that replication attempts generally have, even a true effect will disturbingly often fail to appear successfully replicated if we apply the criterion that Simonsohn (2013) notes is typically employed in the literature: achieving conventional levels of significance.

## The Continuously Cumulating Meta-Analytic (CCMA) Approach

In the lead chapter for the book entitled *Handbook of Replication Research in the Behavioral and Social Sciences*, Rosenthal (1990) critiqued the practice we highlighted above (i.e., determining successful replication based on whether or not the study also achieved significance). He described several more appropriate criteria derived from the meta-analytic framework he helped to popularize. Standard meta-analysis is generally seen as retrospective in nature—a literature review that looks backward to summarize a large set of completed studies. The slight variation we now term *continuously cumulating meta-analysis* (CCMA) performs the exact same meta-analytic calculations but does so in a continuing fashion after each new replication attempt completes. In CCMA, instead of misleadingly noting simply whether each replication attempt did or did not reach significance, we *combine* the data from all the studies completed so far and compute various meta-analytic indexes[3] to index the degree of confidence we can have that a bona fide phenomenon is being investigated. In other words, the individual effect sizes of the entirety of completed studies are pooled into a single estimate. The resulting pooled estimate generally is more trustworthy because it is based on far more data than each individual study. As we discuss below, meta-analysis also allows quantification of the heterogeneity of results. The CCMA approach therefore shifts the question from whether or not a single study provided evidential weight for a phenomenon to the question of how well all studies conducted thus far support conclusions in regards to a phenomenon of interest.

## Simulating Replication Attempts

To illustrate the CCMA framework and how it can outperform a focus on individual studies (and their significance level), we performed a set of simulations.[4] We drew random samples from two normal populations with the identical standard deviations ($\sigma = 1$), but the mean of one population ($\mu_I$) was .5 greater than the other ($\mu_{II}$). Thus, in these populations, Cohen's $d = \frac{\mu_I - \mu_{II}}{\sigma} = 0.5$, which is commonly categorized as a "medium" effect size (Cohen, 1988; $r = 0.243$), and is in the range of effect sizes typically found in published psychological research by recent reviews (Anderson, Lindsay, & Bushman, 1999; Hall, 1998; Lipsey & Wilson, 1993; Meyer et al., 2001; Richard, Bond, & Stokes-Zoota, 2003; Tett, Meyer, & Roese, 1994). We set our cell size at 25 ($n_I = n_{II} = 25$) in each of two groups for a total sample size of $N_1 = 50$),[5] which reflects the median cell size found in recent summaries of extant research (Marszalek, Barber, Kohlhart, & Holmes, 2011; Wetzels et al., 2011). Power calculators, (e.g., http://www.danielsoper.com/statcalc3/calc.aspx?id=49) show that power = .41 for a two-tailed test with these parameters (one-tailed power = .54), which is consistent with the median levels of power that Cohen (1962) and Sedlmeier and Gigerenzer (1989) found in published research.

We drew 10,000 random samples per population group, calculated *t* tests on each, and tabulated the proportion of samples for which the test reached the .05 level. All of our results for the sample sizes and effect sizes above, as well as several other possible sample sizes and effect sizes, are presented in Table 1 (we also list various criteria by which a replication attempt could be evaluated; we discuss each criterion in turn throughout the article).

As a specific example, the results for just the very first of the 10,000 simulated Study 1s are presented in the top row of Table 2.

As shown, this simulated Study 1 yielded a mean difference between the two groups of 0.64 (compared with the population value of 0.5) and a pooled standard deviation of 1.03 (compared with the population 1.0), yielding an effect size estimate of 0.62 (compared with the population value of 0.50), and $t(48) = 2.20$, $p = .033$ (two tailed), which fell below the hallowed .05 level and would therefore be declared significant. Overall, as shown in the first Criterion 1 entry in Table 1, 42% (very close to the power calculator value of .41 noted earlier and listed in the 4th row of Table 1) of the 10,000 samples had results that similarly reached this level. In other words, across 10,000 simulated Study 1s testing this real, medium-sized effect, 42% reached significance.

Then, to simulate a replication attempt with the same sample size *N* (i.e., $N_1 = N_2$), we drew a second set of two samples from the same two populations. As a specific example, examining just the very first of the 10,000 replication attempt samples, whose results are presented in the second row of Table 2, and comparing to the illustrative Study 1 in the line above, we found it yielded a substantially smaller mean difference of 0.40, and a slightly larger pooled standard deviation of 1.07. The effect size estimate here, as a result, was substantially smaller (0.37), as was the *t* test, $t = 1.31$, $p = .198$ (two tailed), which failed to reach the .05 level. Many researchers might regard the second study with the very nonsignificant results as a failure to replicate and would express doubt about the robustness of the effect and perhaps even abandon this line of inquiry. Over the 10,000 samples, as shown in Criterion 2 of Table 1, only 41% of the replication attempt samples reached significance (the value of power again, as Tversky & Kahneman, 1971, noted) and would be declared a successful replication by the "achieve significance" dichotomous standard.

Thus, with typical levels of power and effect sizes, if one uses $p < .05$ as the criterion for a successful replication, both power calculations and our simulations show one will only attain this criterion about 40% of the time when the phenomenon under study is real. Moreover, in only 17% ($\sim.42^2$) of the 10,000 samples were both the original and the replication study significant at .05 (Criterion 3 in Table 1).

Clearly, we expect too much from low power attempts at replication. We could only appropriately use the "achieve significance" dichotomous standard for a successful replication if studies in the field commonly had much higher power: the .80–.95 that so many writers (e.g., Cohen, 1962, 1988; Ellis, 2010) advocate. Until and unless behavioral science research typically proceeds with a much higher degree of power than it currently does, we cannot hope for our discoveries to have the high levels of reproducibility (at conventional levels of significance) that other sciences enjoy. Similar points have been argued by Cumming (2008), for example, who discourages the use of *p* values to indicate replication and champions confidence intervals instead.

In the absence of higher levels of power in typical social science studies, the CCMA perspective we propose recommends an alternative, and more appropriate, criterion that can be used in place of achieving significance to decide the robustness of a phenomenon under study and whether or not a replication was successful (Rosenthal, 1990). A researcher following CCMA procedures, instead of regarding the second study in isolation and making a dichotomous decision about whether it replicated, would combine both studies.[6] Combining the results of the first sample and first replication study in Table 2 yields, as shown, $Z_{\text{overall}} = 2.42$ and an overall *p* value (two-tailed) of .016, which is smaller than that of the first study alone (see Table 2). The effect-size confidence interval, of course, is similarly narrower after the second study's results are combined with the original (not shown). Thus, a CCMA approach would conclude that after both studies were conducted, there is more, not less, evidence that the effect is real. Criterion 4 in Table 1 shows the impact of

**Table 1.** Probability of Significance ($p < .05$) by Various Criteria, for a Range of Study Sample Size ($N$) Cases and Effect Size ($d$) Cases

| Variable | Case 1 | Case 2 | Case 3 | Case 4 | Case 5 | Case 6 | Case 7 | Case 8 | Case 9 |
|---|---|---|---|---|---|---|---|---|---|
| Study 1 $N$ ($N_1$) | 50 | 50 | 80 | 80 | 50 | 50 | 80 | 80 | 50 |
| Study 2 $N$ ($N_2$) | 50 | 80 | 50 | 80 | 50 | 80 | 50 | 80 | 50 |
| $d$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.0 |
| Power (two tailed) according to power calculator for Study 1, given its $N$ | .41 | .41 | .60 | .60 | .10 | .10 | .14 | .14 | 0 |
| Power (two tailed) according to power calculator with combined Study 1 & 2 $N$s | .70 | .79 | .79 | .88 | .17 | .20 | .20 | .24 | 0 |
| Criterion | | | | | | | | | |
| 1. Study 1 achieves .05 | .42 | .41 | .61 | .60 | .10 | .11 | .14 | .15 | .05 |
| 2. Study 2 achieves .05 | .41 | .60 | .42 | .60 | .11 | .14 | .11 | .14 | .05 |
| 3. Studies 1 & 2 both achieve .05 | .17 | .25 | .26 | .36 | .01 | .02 | .01 | .02 | .00 |
| 4. CCMA achieves .05 | .69 | .80 | .81 | .88 | .16 | .19 | .20 | .25 | .05 |
| 5. CCMA achieves .05 and Q-test non significant | .65 | .76 | .77 | .84 | .15 | .18 | .19 | .23 | .05 |
| 6. CCMA achieves .05 and $I^2$ is below 50% | .58 | .67 | .68 | .74 | .13 | .16 | .17 | .21 | .04 |
| 7. Study 2 achieves .05 | .41 | .60 | .42 | .60 | .11 | .15 | .10 | .15 | .04 |
| 8. CCMA achieves .05 | .92 | .95 | .96 | .98 | .62 | .58 | .71 | .68 | .30 |
| 9. CCMA achieves .05 and Q-test non significant | .87 | .90 | .92 | .94 | .58 | .54 | .67 | .64 | .28 |
| 10. CCMA achieves .05 and $I^2$ below 50% | .78 | .79 | .84 | .85 | .49 | .43 | .60 | .56 | .25 |

Note: Criterion 7–10 are all based on the samples in which Study 1 achieves $p < .05$. CCMA = continuously cumulating meta-analysis.

using this superior approach over the 10,000 samples: 69% reached significance (as opposed to only 41% when considering only Study 2's results in isolation.) Note that the second power line of Table 1 shows what a power calculator would show when the $N$ entered is the sum of the two sample sizes: 70%, virtually the same as the CCMA simulated value.

Meta-analysis also provides tests of whether the two studies results differ from one another. This issue of the difference in results was similarly presaged by another of Tversky and Kahneman's (1971) hypotheticals:

> Suppose one of your doctoral students has completed a difficult and time-consuming experiment on 40 animals . . . One comparison yields a highly significant $t = 2.70$, which is surprising and could be of major theoretical significance.
>
> Now suppose the student has repeated the initial study with 20 additional animals, and has obtained an insignificant result in the same direction, $t = 1.24$. What would you recommend now?
>
> Check one (the number of respondents checking each alternative is in parentheses):
>
> (a) He should pool the results and publish his conclusion. (0)
> (b) He should report the results as a tentative finding. (26)
> (c) He should run another group of [median — 20] animals. (21)
> (d) He should try to find an explanation for the difference between the two groups. (30) (pp. 107–108)

As can be seen, not a single one of the Tversky and Kahneman (1971) expert respondents thought that (a) was the best answer.[7] The meta-analytic and CCMA approach, on the other hand, advocates exactly that. But Alternative (d), which was the modal response, suggests we should directly compare (as well as combine) the two studies' results.

Again, human intuition cannot be relied upon, as the experts succumbed once more to belief in the false Law of Small Numbers by perceiving the difference between the two results as sizable and worthy of explanation. Tversky and Kahneman declared "Response *d* is indefensible . . . the difference between the two studies does not even approach significance . . . the attempt to 'find an explanation for the difference between the two groups' is in all probability an exercise in explaining noise" (p. 108).

The CCMA framework again offers the appropriate way to correct fallible intuition and assess quantitatively whether the two studies in Table 2 have obtained results that differ.[8] The question of whether effect sizes are homogenous or heterogeneous can be tested in different ways. Here, we focus on two such methods: The $Q$ statistic, which can be compared against the $\chi^2$ distribution to provide a significance test for effect-size heterogeneity,[9]

**Table 2.** Results From a Simulated Study and a Replication Attempt, With a CCMA Analysis

| Study | Mean diff | $s_{\text{pooled}}$ | $t$ | $p$ | ES (Cohen's $d$) | $Z$ |
|---|---|---|---|---|---|---|
| Original | 0.64 | 1.03 | 2.19 | 0.033 | 0.62 | 2.13 |
| Replication attempt | 0.40 | 1.07 | 1.31 | 0.198 | 0.37 | 1.29 |
| CCMA results | | | | 0.016 | 0.49 | 2.42 |

Note: Homogeneity test was nonsignificant, $Q(1) = .38$, $p = .54$, $I^2 = 0.00$. ES = effect size.

and the $I^2$ measure (Higgins, Thompson, Deeks, & Altman, 2003), which is purely descriptive (rather than involving a significance test). Readers can find formulas and examples for calculating each of these by going to Rosenthal and Rubin (1982) and Higgins et al., respectively. The developers of the $I^2$ index (Higgins et al., 2003) recommend the standard of 25% and 50% for designating small and moderate degrees of heterogeneity, respectively. Patsopoulos, Evangelou, and Ioannidis (2008), consider both of those values. Ioannidis, Patsopoulos, and Evangelou (2007) recommend that confidence intervals around $I^2$ should be routinely computed.

As applied to the illustrative Study 1 and replication attempt portrayed in Table 2, $Q = .38$, $p =.54$, and $I^2 = .0$. Any intuition that suggested the initial results failed to replicate would have to confront the reality that there was barely any statistical difference between the two study's results. Additional results (not shown) indicate that, over all 10,000 samples, only 4.8% of the simulated replication attempts had a significant $Q$ statistic (just about what would be expected, since the null hypothesis of homogeneity was true); moreover, 76% of all the samples had a very small $I^2$ of 25% or less, and 85% had an $I^2$ of less than 50%.

If one had found substantially different effect sizes for the two studies (e.g., a significant $Q$ value and/or a descriptively large $I^2$ value), this result would indicate that homogeneity of effect size is violated, even if the effect sizes are in the same direction—in other words, these indicators would suggest that the two studies yield divergent results. In the context of replication, some would consider this a failure to replicate (Valentine et al., 2011). In any event, it should certainly caution us against interpreting the pooled result as an appropriate summary measure of effect sizes (since the effect sizes were so dissimilar that it is not meaningful to consider a common effect), and spark instead a search for the explanation for the difference between the two groups (i.e., the moderators of effect size).

In contrast, if the second study's results meet the combined criterion that the difference in effect sizes between studies is descriptively quite small and nonsignificant, whereas the pooled result itself is significant, it should increase our confidence that the phenomenon is genuine. The proportion of samples that achieved a significant pooled estimate and had a nonsignificant heterogeneity $Q$ statistic (Criterion 5, Table 1) was 65%, only slightly reduced from the 69% in which heterogeneity was not considered. The proportion of samples that achieved a significant pooled CCMA value and whose descriptive $I^2$ was below 50% was 58%, only slightly lower. We argue that, in the face of the typically low levels of power of current behavioral research, these CCMA values are among the more appropriate indices of the robustness of effects and also give a more appropriate picture of what should count as a successful replication. To aid researchers in obtaining results from a CCMA analysis as described above, we provide supplementary materials (also available online at http://www.human.cornell.edu/hd/qml/software.cfm) in the form of an annotated Excel spreadsheet and a template for R code that uses the "meta" package by Schwarzer (2007). (See the Supplementary Materials section.)

## Attempting Replications Only After the Initial Study Is Significant

A number of writers (e.g., Pashler & Harris, 2012) have pointed out that replication studies are generally undertaken, even by the original researchers, when the original study achieves significance. Initial studies investigating a clearly nonsignificant relationship or phenomenon are shelved (or "file-drawer"-ed; Cooper, 1979; Rosenthal, 1979; see also, http://www.psychfiledrawer.org/about.php), not published, and largely dismissed as false starts.

Thus, if we are to accurately assess how our proposed criteria would fare in the face of this real-world tendency for researchers to only follow up significant Study 1s, we need to investigate the above alternatives as conditional criteria of a successful replication (i.e., conditional on the initial study being significant.) To do so, our simulation set aside those 58% of samples in which Study 1 did not achieve significance, and we studied these various criteria only for the 42% of Study 1s that attained significance. Given a significant effect in Study 1, how often is the effect in Study 2 significant? Of course, this value was within simulation-rounding error of the power calculator value of the test: 41% (see Criterion 7 in Table 1).[10] The probability that the CCMA pooled value over the two studies achieved significance was a reassuring 92% of the samples, as noted by Criterion 8 in Table 1. Likewise, in 87% of the samples in which Study 1 reached $p < .05$, significance was also achieved for the CCMA pooled effect and heterogeneity was nonsignificant (Criterion 9 in Table 1). Finally, in 78% of the samples in which Study 1 reached $p < .05$, significance was achieved for the CCMA pooled effect and $I^2$ was less than 50% (Criterion

10 in Table 1). Thus, after an original study obtains significance, the proper course for the subsequent attempt is to pool its result with Study 1 and look for differences in effect sizes using either significance tests and the *Q* statistic, the descriptive $I^2$ measure, or some other descriptive index of heterogeneity. The bottom line about the Case 1 results in Table 1 (which were based on the typical sample sizes of 25 per group and the typical effect size of 0.50) is that the chances of running another study (with the same *N*) and obtaining a successful replication by this CCMA approach is about 80%.

Table 1 also contains additional columns for other possible sample sizes (80 and 50 for each of the two studies—either 40 or 25 per group) and other effect sizes ($d = 0.2$, commonly deemed a "small" effect size, as well as a null effect of $d = 0.0$; see Cases 8 and 9 of Table 1). As can be seen, small effect-size studies have a disappointing likelihood of achieving significance (generally less than a 20% chance) either individually or even when pooled. Nevertheless, Table 1 shows that if the initial $N = 50$ study does achieve significance (despite the low odds), the pooled result (Criterion 8) has a power of .71 (Case 8) to be significant in the large sample size ($N = 80$, with 40 per group) condition. And even if both the original and the replication attempt study use only 25 subjects per group, the combined pooled criterion is above 60%. True null effects (Case 9 of Table 1), however, tend not to obtain significance by any criteria (other than the conditional ones, due to the fact that the original study was a Type 1 error—in such cases, the cumulative estimate of the effect will move closer and closer to zero as additional studies accumulate, as discussed later in this article).

## Additional Replication Attempts and the Life-Course of Replications

The CCMA approach advocates recomputing meta-analytic indices described in the first section as each new replication attempt is completed. Our confidence should increase that the phenomenon under study is genuine and true to the degree that the combined criterion remains significant, whereas indices of heterogeneity (e.g., *Q* and $I^2$) remain small and nonsignificant. Larger *Q* and $I^2$ values should prompt the search for plausible moderator variables (i.e., dissimilar features of those studies with distinctly differing effect size values).

Figure 1 depicts this outcome (using the combined criterion of significant CCMA and nonsignificant *Q* statistic) over the course of a series of replication attempts, given various effect sizes. These "life-course trajectories" in Figure 1 are for the example of the following sequence of total sample sizes in replication Studies 1 through 7: 50, 50, 20, 70, 70, 50, 80. (We also examined other sample sizes and effect sizes, but figures did not differ appreciably.) Compared with what appears a more-or-less random walk trajectory for the probability that individual studies attain significance, the CCMA approach tends toward consistently higher pooled significance. As can also be seen, the CCMA life course of a true effect is much different than for a null effect (i.e., when $d = 0$). Even for a small effect size, undertaking a series of replication attempts and cumulating their results eventually leads to high probabilities of (combined-pooled) significance. Incidentally, this was exactly the result of a recent replication conducted by the Many Labs project" (Klein et al., 2014). The authors of this orchestrated replication attempt found that of 13 classic findings in psychology (e.g., the anchoring effect), 10 soundly replicated. As an example, the replication attempts of the Quote Attribution study (Lorge & Curtiss, 1936), saw several nonsignificant replications (some even with opposite signs); however, the pooled effect estimate of all replications was extremely narrow and centered around a positive effect of about $d = 0.25$.

In fact, the meta-analytic perspective discourages altogether dichotomous decisions based on .05, even for meta-analytic results such as CCMA tests, and instead advocates employing only continuous descriptive criteria—for example, effect size estimates and their confidence intervals (continuous *p* levels can also be useful, but not dividing them into two categories, the "significant" and the non-significant). This general preference for confidence intervals and effect sizes (and confidence intervals around effect sizes) has been advocated by numerous authors (e.g., Cumming, 2013; Cumming & Finch, 2001; Thompson, 2002). As is well-known, the .05 criterion for what constitutes convincing evidence is completely arbitrary. As Rozeboom (1960) states, "Acceptance of a proposition is not an all-or-none affair; rather it is a matter of degree" (pp. 420–421). In Rosnow and Rosenthal's (1989) words, "surely, God loves the .06 nearly as much as the .05" (p. 1277). Imagine how research reports would change if the word *significant* were stricken from our scientific vocabulary. Researchers would need to argue for each finding's importance anew without relying on rarely applicable one-size-fits-all rules of thumb.

## Recommendations

Our recommendations on how to improve the assessment of evidential weight of numerous studies are based on our understanding of the perfidious impact of belief in the Law of Small Numbers and on the CCMA approach presented here. First, and most important, we need to recognize that the core problem is the very low statistical power of most research in psychological and behavioral science combined with the failure of our intuition to

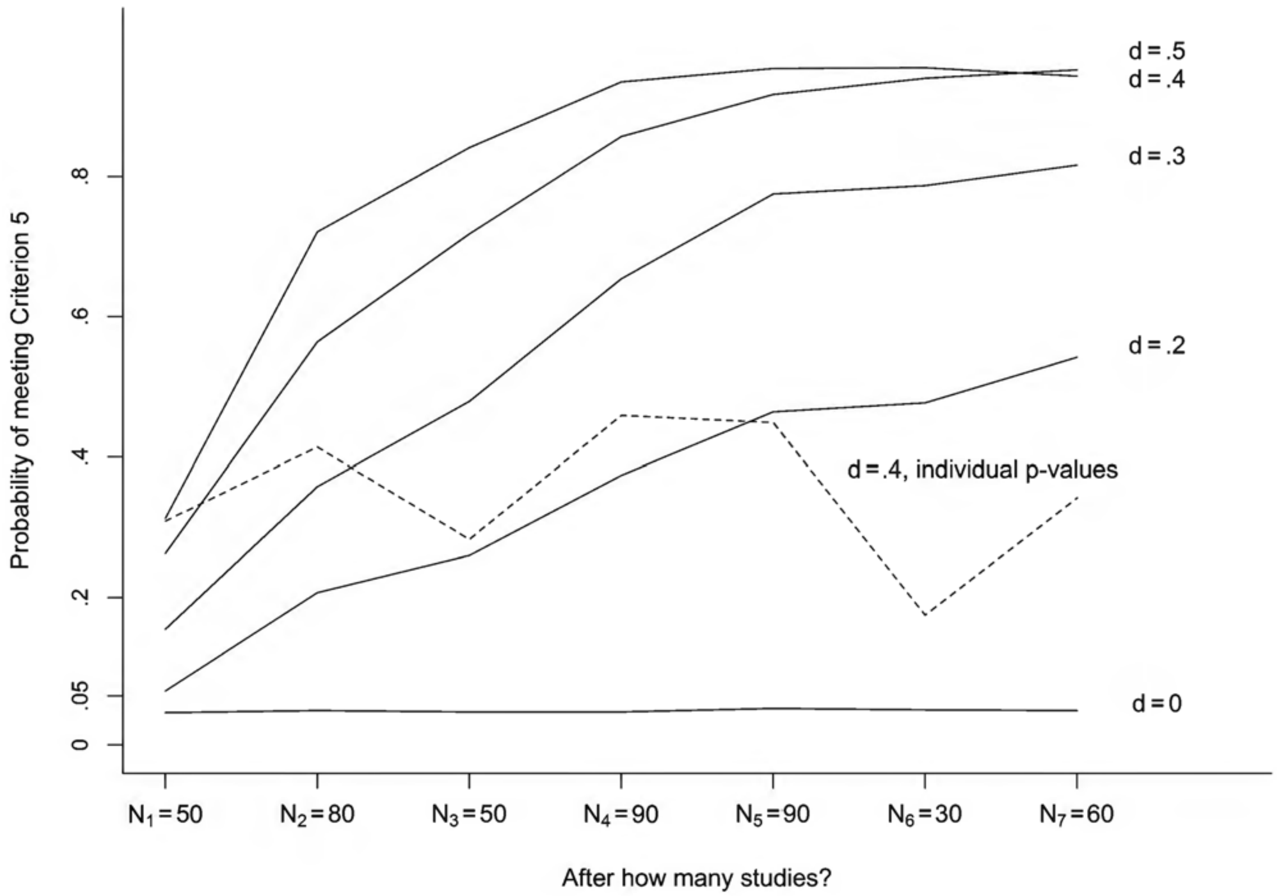

**Fig. 1.** Probability of achieving a significant result for a single study (dashed line) or a significant CCMA estimate (solid lines) using Criterion 5 from Table 1 for a sequence of studies with the sample sizes specified along the *x* axis when testing effect sizes ranging from *d* = 0 to *d* = .5.

anticipate its pernicious effects. The call for higher power is certainly not a new one (see, e.g., Cohen, 1962; Lakens & Evers, 2014, this issue; Sedlmeier & Gigerenzer, 1989), however it is worth repeating. In fact, there are some reassuring signs that this trend is very recently changing (e.g., the Many Labs replication project prioritized adequately powered replication attempts; Perugini, Gallucci, & Costantini, 2014, this issue; Stanley & Spence, 2014, this issue). If all psychological research were carried out with power near .95—or even .80—replication problems would arguably diminish. True effects would virtually always achieve significance and virtually always be replicable, and researchers would start being more interested in effect sizes. We wouldn't have to contend with fallacious beliefs such as the Law of Small Numbers, we'd instead have suitably large numbers. Arguably, *p* hacking (Simmons, Nelson, & Simonsohn, 2011) and other questionable research practices (John, Loewenstein, & Prelec, 2012) would diminish, because significance in high-powered studies would in fact be achieved without tricks. If psychology is serious about solving its inferential problems and achieving the credibility and reproducibility of the physical sciences, this requirement will do the trick (Cohen, 1994).

It is interesting to note that Tversky and Kahneman's (1971) own intuition failed on this point. They wrote "We refuse to believe that a serious investigator will knowingly accept a .50 risk of failing to confirm a valid research hypothesis" (p. 110). On this they were clearly overly optimistic. Virtually all investigators and almost all publication outlets have contributed to the field being overrun with very underpowered studies despite decades of knowing of the problem.

In any event, we need a better lens with which to view replication efforts than whether they achieve significance. The field has come to accept meta-analysis as the standard for conducting retrospective literature reviews. We need to adopt a similar perspective, a continuously cumulating meta-analysis (CCMA), to evaluate the validity of research results as each replication attempt is obtained. We have demonstrated that such an approach yields not only a more accurate, but also a richer picture of the pooled effect of numerous studies. The pooled effect estimates are a mathematically sound way to combine effects and indices of heterogeneity help to quantify the amount of discrepancies among studies.

The CCMA approach however is no panacea to fix each and every problem in the field. If studies in the literature have been *p* hacked and thereby overestimate effect sizes, this will bias the CCMA estimate as well. Likewise, publication bias (and the omission of negative results in the literature) can influence CCMA estimates. However, indices of heterogeneity and measures like the funnel plot (Sterne & Egger, 2001) can help to identify publication bias. In addition, alternative measures such as *p* curves (Lakens, 2014; Nelson, Simonsohn, & Simmons, 2014; Schimmack, 2012) can help to identify if studies have been severely *p* hacked and such studies could be excluded from a CCMA. Using and encouraging others to use CCMA could also help decrease *p* hacking and publication bias: If a nonsignificant Study 2 can be published as part of a package of studies that together produce a significant CCMA, researchers will be more likely to include nonsignificant individual findings in their papers (see also Maner, 2014, this issue).

However, the CCMA approach need not be confined to replication studies, but can likewise be used to combine internal replications of multistudy articles. This would be far more informative then simply reporting whether each single study succeeded or failed. We have reviewed the statistical tools to conduct CCMA analyses and hope that readers will implement these tools in their own research and encourage others to do so when reviewing articles or replication proposals.

## Acknowledgments

The authors are grateful to the following readers of earlier drafts for helpful comments: Todd S. Braver, David Funder, Robert J. MacCoun, Irwin N. Sandler, and Stephen G. West. We also benefitted from discussions with Deanna M. Barch, Daniel Ozer, Piotr Winkielman, and the participants in the Quantitative Seminar at University of California, Riverside.

## 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] Tversky passed away before the Nobel prize was given to Kahneman, but Kahneman (2002) acknowledged they were essentially joint winners.

[2] A fifth determinant is the properties of the exact statistical procedure conducted.

[3] A huge literature exists on meta-analytic techniques. Good readable summaries are Lipsey and Wilson (2001) and Rosenthal (1984). Many reasonable meta-analytic indices have been proposed, and we don't mean to imply by our choices of ones to feature here that these are necessarily the superior ones. The indices we feature are $Z_{overall}$, which we refer to as CCMA (for formulas and discussion, see Rosenthal, 1984, 1990; Rosenthal & Rubin, 1979, 1982); Q (Hedges, 1982; Rosenthal, 1984; Rosenthal & Rubin, 1982); and $I^2$ (Higgins et al., 2003).

[4] Many of the results presented herein can be analytically derived, but we used simulations throughout for consistency and to make the points concrete.

[5] Our notation uses lower case *n* with Roman numeral subscripts to refer to sample size per group (generally we consider only two groups of equal size, so $n_I = n_{II}$), and capital *N* with numeric subscripts to refer to sample size per study.

[6] The easiest way of doing so is by converting each of the one-tailed *p* values to a *Z*. The Excel function NORMSINV is one way of obtaining these values. For example, consider the second row of Table 2, in which $t = 1.31$ and $p = .198$. The .198 is the two-tailed *p* value. To obtain the *Z* for the one-tailed *p* value (which is shown to equal 1.29), use "=NORMSINV(1−.198/2)". Then the various resulting *Z* values are summed and the total is divided by the square-root of the number of *Z*s (or studies), in this case $\sqrt{2}$. Alternative meta-analytic formulas (for example, those given by Borenstein, Hedges, Higgins, & Rothstein, 2011) are also available for computing a significance test of the pooled effect estimate. These meta-analytic pooled estimates can be based on either fixed or random-effects models. If assuming direct replications, a fixed-effect model is appropriate. Our results in Tables 1 and 2 are based on fixed-effects models. Random-effects models have slightly larger standard errors and thus wider confidence intervals. We obtained results for random-effects models as well, but other than being slightly less powerful, patterns of results did not change.

[7] Tversky and Kahneman (1971) actually added the words "as fact" at the end of the phrase for Alternative (a), which is perhaps why not a single respondent chose it. They also thought that Alternatives (b) and (c) were acceptable and could "be justified on some grounds" (p. 108).

[8] The CCMA approach produces the same mean estimate of the cumulative effect size as a standard Bayesian approach but adds the investigation of homogeneity. On the other hand, it loses information about the distribution of the posterior, which might also be quite interesting.

[9] These indices weight large *N* studies more highly than low *N* studies, based on the recognition that their respective effect-size estimates are differentially precise. The statistical power of the Q test has been frequently recognized as rather low (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006), in that it fails to detect as significant even sizable differences in effect size. This poses an especial problem for tests of homogeneity because the researcher commonly hopes not to reject the null hypothesis. Some (e.g., Berlin, Laird, Sacks, & Chalmers, 1989; Fleiss, 1993; Petitti, 2001) recommend raising the power of such tests by raising their alpha to .10 rather than .05.

[10] The homogeneity of the effect size *Q*-statistics being significant remained quite similar at 5.6% (not shown).

## References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory. *Current Directions in Psychological Science*, *8*, 3–9. doi:10.1111/1467-8721.00002

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Bartlett, T. (2012). Is psychology about to come undone? *Chronicle of Higher Education*. Retrieved from http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045

Berlin, J. A., Laird, N. M., Sacks, H. S., & Chalmers, T. C. (1989). A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine*, *8*, 141–151.

Bernoulli, J. (1713). *Ars conjectandi: Usum & applicationem praecedentis doctrinae in civilibus, moralibus & oeconomicis*. (O. Sheynin, Trans.). Berlin, Germany: Verlag.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Available from Wiley.com

Carpenter, S. (2012). Psychology's bold initiative: In an unusual attempt at scientific self-examination, psychology researchers are scrutinizing their field's reproducibility. *Science*, *335*, 1558–1560.

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

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

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997–1003.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, *37*, 131–146.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, England: Routledge.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York City, NY: Cambridge University Press.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, *2*, 121–145.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. doi:10.3758/s13423-012-0227-9

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571.

Hall, J. A. (1998). How big are nonverbal sex differences? In D. J. Canary & K. Dindia (Eds.), *Sex differences and similarities in communication* (pp. 155–177). Mahwah, NJ: Erlbaum.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, *8*(8), e72467.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490–499.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I 2 index? *Psychological Methods*, *11*, 193–206.

Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal*, *335*, 914–916.

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.

Kahneman, D. (2002). *Prize lecture: Maps of bounded rationality*. Retrieved from http://www.nobelprize.org/nobel_prizes/economics/laureates/2002/kahneman-lecture.html

Klein, R., Ratliff, K., Vianello, M., Adams, R., Bahink, S., Bernstein, M., . . . Nosek, B. (2014). *Investigating variation in replicability: A "Many Labs" Replication Project*. Retrieved from https://osf.io/wx7ck/

Lakens, D. (2014). *Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses*. Retrieved from http://dx.doi.org/10.2139/ssrn.2381936

Lakens, D., & Evers, E. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives in Psychological Science*, *9*, 278–292.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: SAGE.

Lorge, I., & Curtiss, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology*, *7*, 386–402.

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

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives in Psychological Science*, *9*, 343–351.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, *112*, 331–348. doi:10.2466/03.11.pms.112.2.331-348

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–156. doi:10.1037/0003-

Nelson, L., Simonsohn, U., & Simmons, J. (2014). *P-curve fixes publication bias: Obtaining unbiased effect size estimates from published studies alone*. Retrieved from http://dx.doi.org/10.2139/ssrn.2377290

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. In J. W. Neuliep (Ed.), *Handbook of replication research in the behavioral and social sciences* (pp. 85–90). Corta Madera, CA: Select Press.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality*, *8*(6), 21–29.

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

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

Patsopoulos, N. A., Evangelou, E., & Ioannidis, J. P. (2008). Sensitivity of between-study heterogeneity in meta-analysis: Proposed metrics and empirical evaluation. *International Journal of Epidemiology*, *37*, 1148–1157.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives in Psychological Science*, *9*, 319–332.

Petitti, D. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine*, *20*, 3625–3633.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–664.

Rosenthal, R. (1984). *Meta-analytic procedures for social sciences*. Beverly Hills, CA: Sage.

Rosenthal, R. (1990). Replication in behavioral research. In J. W. Neulip (Ed.), *Handbook of replication research in the behavioral and social sciences* (pp. 1–30). Corta Madera, CA: Select Press.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, *56*, 1165–1168.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Rosnow, R. L., & Rosenthal., R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

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

Schwarzer, G. (2007). Meta: An R package for meta-analysis. *R News, 7*(3), 40–45.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

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.

Simonsohn, U. (2013, December 10). *Small telescopes: Detectability and the evaluation of replication results*. Retrieved from http://dx.doi.org/10.2139/ssrn.2259879

Stanley, D., & Spence, J. (2014). Expectations for replications: Are yours realistic? *Perspectives in Psychological Science*, *9*, 305–318.

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*, 1046–1055.

Tett, R. P., Meyer, J. P., & Roese, N. J. (1994). Applications of meta analysis: 1987–1992. *International Review of Industrial and Organizational Psychology*, *9*, 71–112.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. doi:10.1037/h0031322

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., ... Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*, 103–117.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

Yong, E. (2012, March). A failed replication attempt draws a scathing personal attack from a psychology professor. *Discover Magazine*. Retrieved from http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/#.U0gDgPldVc4

ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

# Erratum

Perspectives on Psychological Science
2014, Vol. 9(4) 452
© The Author(s)
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614542784
pps.sagepub.com
SAGE

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342. (Original doi: 10.1177/1745691614529796).

The Supplemental Material paragraph was missing from the text. It appears below:

**Supplemental Material**

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data

# Erratum

Perspectives on Psychological Science
2014, Vol. 9(4) 452
© The Author(s)
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614542786
pps.sagepub.com
SAGE

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*, 319–332. (Original doi: 10.1177/1745691614528519).

The Supplemental Material paragraph was missing from the text. It appears below:

**Supplemental Material**

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data

# Erratum

Perspectives on Psychological Science
2014, Vol. 9(4) 452
© The Author(s)
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614542787
pps.sagepub.com
SAGE

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9*, 305–318. (Original doi: 10.1177/1745691614528518).

The Supplemental Material paragraph was missing from the text. It appears below:

**Supplemental Material**

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data