Meta-analysis to integrate effect sizes within a paper: possible misuse and Type-1 error inflation

Article

Accepted Version

Running head: Meta-analysis & false positives

Meta-analysis to integrate effect sizes within a paper: Possible misuse and Type-1 error inflation

Taiji UENO[1,2,*]

Greta M. FASTRICH[1]

Kou MURAYAMA[1,3*]

[1]Department of Psychology, University of Reading, UK

[2] Department of Psychology, Nagoya University, Japan

[3]Kochi University of Technology

* Correspondence to:
Dr. Taiji Ueno & Dr. Kou Murayama

[Taiji Ueno]
Department of Psychology, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
Tel: +81-52(789)4716
Email: ueno.taiji@d.mbox.nagoya-u.ac.jp
[Kou Murayama]
2S23, Department of Psychology, University of Reading, Earley Gate, Whiteknights, Reading RG6 6AL, UK.
Tel: +44-(0)118-378-5558 (FAX 6715)
Email: kou.murayama@reading.ac.uk

Conflicts of interested: None declared.
We would like to thank Ms. Ayaka Tokai (Nagoya University) for collecting the papers to review. Taiji Ueno is a JSPS Research Fellow. This work was supported by JSPS KAKENHI Grant Number (14J02696) to TU and Grant Number (15H05401) to KM, and the Marie Curie Integration Grant (CIG630680) by European Commission to KM.

KM and TU generated the research design. TU conducted literature review. GF conducted a simulation under the supervision of KM. TU generated the first draft, and everyone involved the discussion and writing up.

Abstract

In recent years an increasing number of papers have employed meta-analysis to integrate effect sizes of researchers' own series of studies within a single paper ("internal meta-analysis"). Although this approach has the obvious advantage of obtaining narrower confidence intervals, we show that it could inadvertently inflate false-positive rates if researchers are motivated to use internal meta-analysis in order to obtain a significant overall effect. Specifically, if one decides whether to stop or continue a further replication experiment depending on the significance of the results in an internal meta-analysis, false-positive rates would increase beyond the nominal level. We conducted a set of Monte-Carlo simulations to demonstrate our argument, and provided a literature review to gauge awareness and prevalence of this issue. Furthermore, we made several recommendations when using internal meta-analysis to make a judgment on statistical significance.

Key words: meta-analysis, false positives, replications

The prevalence of underpowered studies in psychology has been repeatedly remarked on in the past (e.g., Cohen, 1988; Sedlmeier & Gigerenzer, 1989), but it is only recently that researchers have started to seriously consider the problems and implications of underpowered research (e.g., Button et al., 2013). This increased attention is obviously in part caused by the recent "replication crisis" in psychology. For example, Ioannidis (2005) showed that underpowered studies are problematic, not only because these studies are unlikely to discover true effects (false negatives), but also because the prevalence of underpowered research actually lowers the possibility that statistically significant results reflect a true effect. In addition, underpowered studies are susceptible to questionable research practices, such as flexibility in analytic choices (Rosenthal, 1979; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014), which are likely to increase false-positive findings. In fact, a recent large-scale reproducibility project (Open Science Collaboration, 2015) found that the effect sizes in (high powered) replication studies were half the magnitude of the original effects of 100 articles in high-profile journals from cognitive and social psychology (see also, Klein et al., 2014; Open Science Collaboration, 2012, 2013; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012). LeBel (2015) also argued that this issue is not limited to cognitive and social psychology, but is prevalent in most fields of psychology (or perhaps in other scientific fields such as neuroscience). These findings clearly demonstrate the value of conducting high powered research in psychology to accurately ascertain effect size.

A variety of measures have been proposed to address this issue (e.g., high-powered study pre-registration; Chambers, 2013), and these initiatives have gradually changed how research is conducted and reported. Among them, one notable change in the recent literature of psychology is the more frequent use of meta-analyses to integrate research findings from multiple studies

within a single paper (hereafter we shall call this *internal meta-analysis*). Commonly, meta-analysis is used to synthesize the findings across the literature from different authors (e.g., Glass, 1976; see also, Open Science Collaboration, 2015, for the meta-analysis of the original effect size and the replication effect size), but in contrast to this, internal meta-analysis aims to integrate researchers' own studies (typically a handful of studies). For example, psychologists often conduct two or more studies in which their key hypothesis is repeatedly tested either with slight modifications (i.e., conceptual replication) or with exactly the same methods but for different samples (i.e., direct replication). Then the effect sizes of the tested-hypothesis in each study are integrated by a meta-analysis in order to draw robust statistical inferences with a larger sample size. Internal meta-analysis has attracted increasing attention since Geoff Cumming's seminal work (2008, 2012, 2013) explicitly encouraged psychologists to conduct meta-analysis on authors' own multiple replication studies to provide a more precise estimation (i.e., narrower CI) of the population effect size (see also, Braver, Thoemmes, & Rosenthal, 2014). Cumming also released supplementary software, Exploratory Software for Confidence Intervals (ESCI) in his book (Cumming, 2012), which has provided an excellent platform for psychologists to conduct meta-analysis.

Potential Inflation of False-positive Rates in Internal Meta-analysis

Meta-analysis is undoubtedly useful to obtain a more precise estimation of the population effect size and increase statistical power by combining multiple studies (Alogna et al., 2014; Donnellan, Lucas, & Cesario, 2015). Accordingly, internal meta-analysis should provide a good potential solution to overcome the lack of statistical power frequently observed in psychology studies. This is indeed true if this analytic method is *appropriately used*. What has been little recognized in the literature, however, is the fact that internal meta-analysis can inflate false

positive error rates (Type-1 error rates) if researchers are not aware of potential issues that often arise in the research planning process. Imagine conducting two studies (on the same phenomenon) with one obtaining a non-significant effect and the other a significant effect. You then conduct a meta-analysis of these two studies but the integrated result is not entirely clear; the integrated effect is close, but not statistically significant. After this, you decide to conduct one more replication study, and the integrated results turn out to be statistically significant. You thus publish a paper in favor of your hypothesis with these three studies. Situations like this could happen in practice, but this practice actually increases the rate of false-positive findings. The critical problem is making a decision to continue or stop conducting a further study *after* looking at the results of an internal meta-analysis. If researchers are motivated to arbitrarily stop replication attempts/studies when the integrated effect becomes significant ($p < .05$) in an internal meta-analysis, false-positive rates will increase.

This issue of the *flexible stopping rule* is not very new. For example, Simmons et al. (2011) argued that flexibly increasing the sample size within a single study until statistical test reveals a significant effect (e.g., $p < .05$) would increase false positive rates, sometimes to a considerable degree. Similar problems and potential solutions have also been noted especially in clinical research. For example, in a clinical trial (i.e., a single large-scale study), researchers often recruit a group of participants multiple times, and conduct an interim analysis after each recruitment to statistically test the effect of treatment (mainly with the aim to finish the trial as early as possible to prevent unnecessary exposure of patients to a potentially unsafe treatment). Several statistical approaches have been offered in response to the potential inflation of Type-1 error rates that such practice can result in, such as an adjustment of the nominal level (group-sequential tests, Lan & DeMets, 1983; O'Brien & Fleming, 1979; Pocock, 1977; for other

alternatives, Lehmacher & Wassmer, 1999; Müller & Schafer, 2001). In another line of research of clinical trials, researchers synthesize the results from a new study with those from the previous literature every time a new study is published (see, Berkey, Mosteller, Lau, & Antman, 1996; Lau et al., 1992). As this "cumulative meta-analysis" also involves sequential statistical testing, it would inflate Type-1 error rates if researchers analyzed the data in the same way as a standard meta-analysis. However, researchers have been well aware of the issue since the pioneering work of Pogue and Yusuf (1997) and several formal solutions to prevent Type-1 error inflations in cumulative meta-analysis have been proposed (Brok, Thorlund, Gluud, & Wetterslev, 2008; Higgins, Whitehead, & Simmonds, 2011; Wetterslev, Thorlund, Brok, & Gluud, 2008). One popular solution is trials sequential analysis (TSA), where researchers evaluate the significance of the integrated effect size not only by the nominal statistical significance of the test, but also by the statistical power and the magnitude of the test statistics (TSA: Brok et al., 2008; Wetterslev et al., 2008; for extensions of TSA, see, Miladinovic et al., 2013; Thorlund et al., 2009). Several other different ways to prevent inflation of Type-1 error rates have been discussed in the literature (e.g., semi-Bayesian approach, Higgins et al., 2011; recalculation of sample size after each cumulative meta-analysis, Roloff, Higgins, & Sutton, 2013; use of triangular boundaries, van der Tweel & Bollen, 2010).

Despite past literature addressing flexible stopping rule and how it inflates Type-1 errors, this issue has never been discussed in the context of internal meta-analysis. We believe it is particularly important to raise the issue in the context of internal meta-analysis for four reasons. First, in comparison to cumulative meta-analysis, internal meta-analysis synthesizes the studies from a single group of authors, not from the past literature. This means that the analysis process is less transparent to readers in internal analysis than in cumulative meta-analysis, providing the

authors more opportunities to "game the system" to gain statistical significance. In fact, as discussed later, we found that a number of studies failed to report sufficient information to replicate the internal meta-analysis results. Second, unlike clinical trials where a third party is typically involved in monitoring the data acquisition process , it is typically easier to collect additional data (even for a new experiment) in psychology, indicating that the potential impact of flexible stopping rule in internal meta-analysis may be more serious than researchers commonly expect (as illustrated by Simmons et al., 2011). This point is particularly relevant to internal meta-analysis given that a number of editors/journals have begun to adopt new policy to exclude flexible stopping rules at the participant-wise level (e.g., asking authors to declare how they determined the sample size for each study) --- it is possible that flexible stopping rule in internal meta-analysis could be used as the next loophole. Third, researchers commonly consider it appropriate or methodologically rigorous to conduct another study to examine the robustness of their findings. In fact, this replication attempt is good practice in itself (e.g., Braver et al., 2014), and our point is that it can inflate Type-1 error rates only when it is combined with internal meta-analysis. However, given that replications are indubitably a good scientific practice, it is possible that researchers are not explicitly aware of the flexible stopping rule they may be using when conducting internal meta-analysis. Fourth, given the general skepticism about replicability of psychological research in recent years, meta-analysis is becoming a more and more popular tool to evaluate the replicability of the previous findings (Klein et al., 2014; Open Science Collaboration, 2015). We can foresee that increased numbers of papers would use internal meta-analysis to demonstrate a replicability of their own findings. Thus, illustrating the potential misuse of internal meta-analysis is both timely and important to future research practice.

Two notes should be made about our arguments. First, our argument does not contradict Cumming's original spirit of internal meta-analysis. Specifically, the potential inflation of Type-1 error rates in internal meta-analysis hinges on researchers' motivation (be it explicit or implicit) to make a dichotomous decision about statistical significance (i.e., "Is the overall effect statistically significant?"). Cumming (2008, 2012, 2013) has also explicitly rejected reliance on null-hypothesis significance testing (NHST). He argues in his papers for a focus on estimation (confidence intervals: CI), rather than a dichotomous decision. In other words, when he introduced the idea of internal meta-analysis, his focus was to distillate a single point estimate with a narrower CI to obtain a more accurate effect size estimate but he never recommended researchers to test statistical significance of the overall effect. Indeed, Cumming (2012) explicitly emphasized that a NHST and a reference to a *p*-value should be abandoned throughout his book about statistics. Of course, as Trafimow and Marks (2015) argue, users could still draw a dichotomous decision from the estimated CI, and this is unfortunately the case in current research practice (see also, Morey, Rouder, Verhagen, & Wagenmakers, 2014). As long as researchers have a dichotomous decision to make on significance, our point is still relevant even if CIs are used. Second, the flexible stopping rule is a separate problem from another well-known major problem in meta-analysis --- publication bias. Publication bias is the tendency on the part of investigators or editors to be reluctant to publish null findings. Not reporting studies with null findings typically leads to the overestimation of the population effect size (Egger et al., 1997), which results in the inflation of false-positive rates. This reporting bias should also be the case for internal meta-analysis, but our argument indicates the potential inflation of Type-1 error rates even in situations where researchers reported all the studies they conducted (but see our Simulation 2 for a more nuanced view on this issue).

In the following section, we will first briefly describe the general procedure of a meta-analysis to guide our research questions. We will then set out Monte-Carlo simulations that empirically demonstrate how the flexible stopping rule in an internal meta-analysis could inflate false positive rates. Finally, we will provide a literature review of the published articles in psychology research that made use of internal meta-analysis to gauge the awareness and possible prevalence of the issue.

Meta-analysis

This section will describe the general statistical framework for conducting meta-analyses. The first step is to estimate an effect size and its sample variance in each study. Depending on the variable types (e.g., binary, continuous, etc.), we use different formulas to compute effect size such as Cohen's *d* or relative risk ratio. Readers are referred to the Appendices for details. The next step is to integrate the effect sizes from all the studies to estimate the single overall effect size and its variance. There are two classes of models to integrate effect sizes: One is the *fixed-effect model* and the other is the *random-effects model*. In fixed-effect model, the true (population) effect size in each study is assumed to be identical across studies. If two observed effect sizes differ between studies, the difference should be entirely due to the sampling variation of participants for each study. In other words, if you increase the number of participants up to infinity for each study, all the observed effect sizes should converge to the same true population effect size. In random-effects model, on the other hand, the true effect size is assumed to vary across studies depending on study-specific factors (e.g., ages, countries, items, etc.; see, Borenstein, Hedges, Higgins, & Rothstein, 2010, for the details of these differences). Under this model, if you increase the number of participants up to infinity in each study, the observed effect sizes should still vary, as the true effect sizes are different between studies. Typically, the true

effect size for each study is posited to follow a normal distribution, N (μ, τ^2). τ^2 represents the between-studies sampling variance of the population effect sizes. Thus, τ^2 is assumed to be 0 in the fixed-effect model; in other words, fixed-effect model is a special case of random-effects model.

There is no explicit guideline on which procedure would be the best for internal meta-analysis. When integrating effect sizes from different papers, some researchers may want to say that all the included studies are similar enough to safely assume a priori that τ^2 is negligible, especially in the case of direct replication. If this assumption is correct, fixed-effect model would have higher statistical power (Hedges & Olkin, 1985). However, even methodologically similar studies can often produce heterogeneous results for various reasons (including direct replications, see, Klein et al., 2014; Open Science Collaboration, 2015). In addition, even if τ^2 is absent, random-effects model would be more likely to produce a non-significant τ^2 estimate. Thus, random-effects model may generally be a safe option for internal meta-analysis unless we are strongly confident that τ^2 is almost negligible. That said, when τ^2 is large and the number of studies is small, random-effects model is known to increase Type-1 error rates (IntHout, Ioannidis, & Borm, 2014; see Figure S1, for example); we need to be careful about this limitation of the random-effects model.

The point estimate of the integrated effect size (*ES*) and its variance (*V*) are computed in the following equations:

$$ES = \frac{\sum_{j=1}^{n}\left(W_j * ES_j\right)}{\sum_{j=1}^{n} W_j}$$

$$V = \frac{1}{\sum_{j=1}^{n} W_j}$$

where subscripts refer to the study ID, ES_j refers to the effect size of the *j*-th study, and W_j refers to the weight for the effect size of the *j*-th study which is computed as:

$$W_j = \frac{1}{\left(\sigma_j\right)^2 + \tau^2}$$

where σ_j refers to the sampling variance of the effect size of the *j*-th study (ES_j). Again, in a fixed-effect model, $\tau^2 = 0$, whereas $\tau^2 \neq 0$ in a random-effects model. Of course, the population parameter τ^2 is unknown. There are several different ways to estimate τ^2, but one common way is to use the method-of-moments estimator T^2 which is calculated by the following formula (DerSimonian & Laird, 1986).

$$T^2 = \frac{Q - df}{\sum_{j=1}^{n} W_j - \frac{\sum_{j=1}^{n}\left(W_j\right)^2}{\sum_{j=1}^{n} W_j}}$$

$$df = \text{number of studies to be integrated} - 1$$

where Q refers to the heterogeneity (variability) between-studies effect sizes, and *df* is the number of studies being integrated minus one.

,

$$Q = \sum_{j=1}^{n}\left[W_j * \left(ES_j - ES\right)^2\right]$$

Once the single point estimate of the effect size (ES) and its variance (V) are calculated, these estimates are used to estimate the C% confidence interval of the population effect size, by consulting the z-distribution.

$$\text{C\% CI} = \left[ES\ -\ Z_{c/100} * \sqrt{V}\ ,\qquad ES + \ Z_{c/100} * \sqrt{V}\right]$$

As can be seen from the formula, a larger number of studies would normally produce narrower CIs (but not always so for a random-effects model meta-analysis in the face of heterogeneity, see: Cohn & Becker, 2003), thereby allowing a more precise estimation of the population parameter (Cumming, 2008, 2012, 2013). Although not recommended by Cumming (2008), it is possible to perform a conventional NHST, using this estimated 95% CI. Specifically, if the 95% CI did not contain 0, then we can argue that the effect size is statistically significant ($p < .05$). Therefore, a narrower CI assures higher detection power of a statistically significant effect.

Monte-Carlo Simulation of False Positive Rates in Internal Meta-analysis

Simulation 1: Flexible stopping rule and false-positive rates

To illustrate our points, we set out a series of Monte-Carlo simulations to examine how the flexible stopping rule would inflate Type-1 error rates in internal meta-analysis. Again, flexible stopping rule in this article is defined as the decision not to conduct any more replication studies upon obtaining a significant result. We simulated a series of studies with two independent groups (N = 20 for each group) within a paper and statistically tested the mean differences between the groups. For each study, the data for the two groups were generated from the standard normal distributions, N (0, 1); therefore, all the statistically significant effects in the

simulations can be deemed false-positives. Note that this data generation procedure assumes a situation where the fixed-effect model is correct, as the true mean difference is always zero in all the simulated studies (i.e., no between-study variation in the true effect size). As our literature reviews shows later (Table 2), heterogeneity (*Q*-index and *I*-squared) across the integrated studies was generally small in the papers that used internal meta-analysis. This fact motivated us to set a simulation of fixed-effect model for our primary simulations (see Appendices for simulations with random-effects model). Note, however, that we do not intend to argue that a fixed-effect model should always be the first choice to analyze data. As mentioned above, even methodologically similar studies can yield very different results (Borenstein, Hedges, Higgins, & Rothstein, 2009). The statistical tests were conducted as follows: for the 1st study, an independent-samples *t*-test was conducted because it is not possible to run a conventional meta-analysis on a single study. From the 2nd study onwards, a meta-analysis was applied after every study to integrate all the studies performed at that point (we used an R package, *meta*: Schwarzer, 2012).

Our primary purpose was to examine the false-positive rates if a researcher uses a flexible stopping rule --- researcher stops conducting further studies once an internal meta-analysis produced a statistically significant effect (when statistical analyses returned a *p*-value of lower than 5% for the first time). Figure 1 shows the simulation results (replication = 15,000), plotting false-positive rates against the maximum number of studies (*j*) that a researcher intended to conduct (i.e., the probability of obtaining significant effects until the researcher gets to the *j*th study or obtaining significant effects at the *j*th study). We compared the scenarios where 1) a researcher consistently applied fixed-effect model meta-analysis, 2) a researcher consistently applied random-effects model meta-analysis, and 3) a researcher simply pooled the raw data and

conducted independent samples *t*-tests, rather than conducting a meta-analysis. Pooling the raw data from different studies is not generally recommended unless researchers take into account the clustering of participant data within studies (Cooper & Patall, 2009; Riley, Lambert, & Abo-Zaid, 2010), but we included this for the purpose of comparison. Indeed, one of the papers which we review in the next section conducted this type of analysis (without accounting for clustering within studies). To facilitate the comparison, in the current simulation, all the data are generated from the same distribution, assuming no clustering within studies (i.e., the potential problem of pooling raw data is underestimated). Finally, we also simulated the situation where internal meta-analysis was appropriately used. That is, we simulated the situation where a researcher simply ran *j* studies and conducted a meta-analysis to integrate these *j* results to test statistical significance.

As expected (Figure 1), false positive rates were 5% for the 1st study (independent samples *t*-test) in all the instances, and remained the same after the second study when internal meta-analysis was appropriately conducted (i.e., no flexible stopping rule was used). In contrast, when researchers were motivated to make a decision to continue or stop further studies, Type-1 error rates monotonically increased as the maximum number of intended studies increased, irrespective of the statistical tests used. It may be unlikely that researchers would run up to eight additional studies in reality; the simulations for such cases were conducted just for illustrative purpose. These findings illustrate the potential problem of internal meta-analysis when researchers make a decision to continue or stop further studies based on the obtained results. We also observed that random-effects model meta-analysis showed slightly lower Type-1 error rates than fixed-effect model meta-analyses, even though we generated data from a fixed-effect model (i.e, a use of fixed-effect model meta-analysis was correct). One likely reason is that estimates of

heterogeneity are sometimes non-zero by chance even if data were generated from the population without heterogeneity, leading to wider confidence intervals.

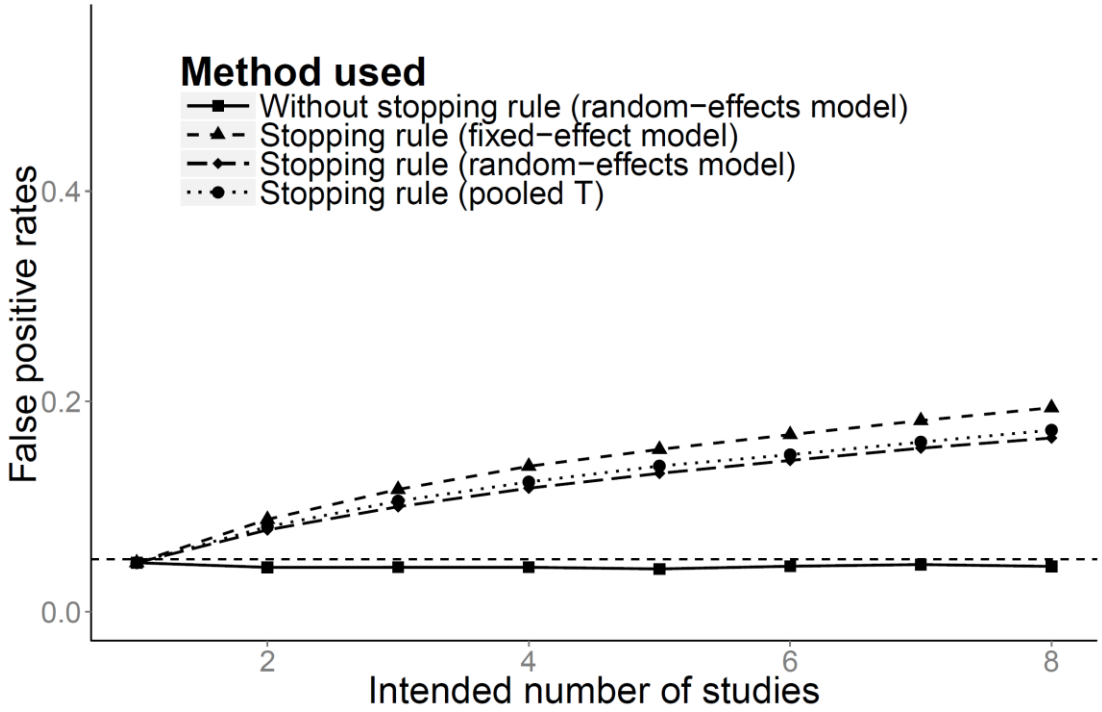


Figure1. Likelihood of obtaining a false positive result when the flexible stopping rule is used (i.e., researchers stop running a new study upon obtaining significance, $p < .05$) in a fixed-effect model meta-analysis (triangle marker), a random-effects model meta-analysis (rhombus marker), and a pooled t-test (circle marker) as a function of the number of studies that the researcher intended to conduct at the maximum. The analysis methods are written in the parentheses. Square makers refer to the likelihood of a false-positive result without the flexible stopping rule (researchers run studies until the intended number of studies are completed irrespective of statistical significance) in a fixed-effect meta-analysis. Note that an independent samples t-test was conducted when the 1st study was conducted (see main text).

We also conducted additional Monte-Carlo simulations by changing the number of participants. Increasing the number of participants for each study (N = 10, 20, or 50 per group)

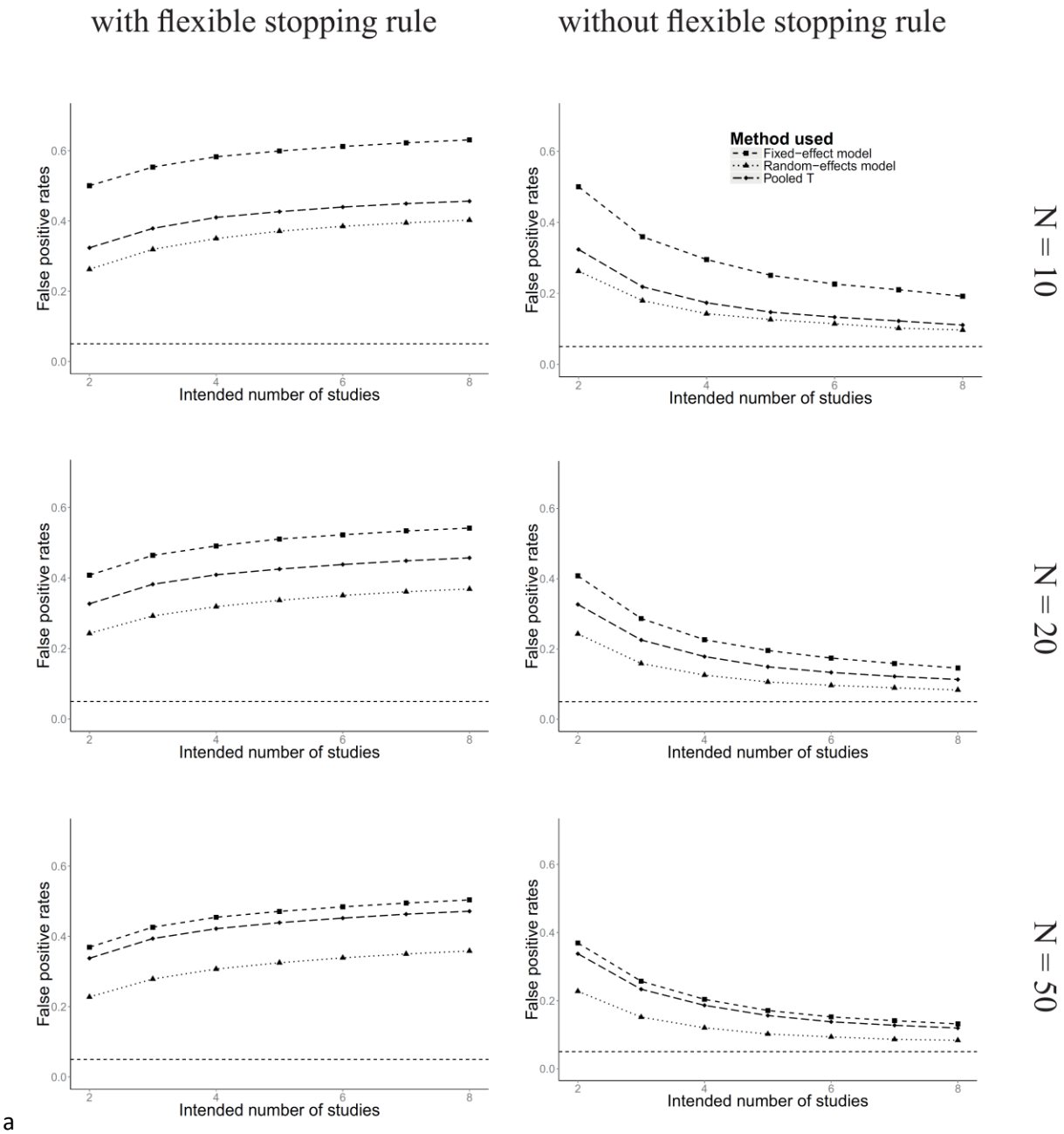
generally decreased the false-positive rates, but did not have a big influence on the pattern of results (see left column of Figure S1 in Appendices). These results indicate that increasing the number of participants for each study partially addresses the issue, but did not solve the problem.

The between-studies variance can be non-zero even if the design involves direct replications (though it seems to be small in the cases of our reviewed papers – see Table 2). As such, we repeated the same set of simulations with data generated from a random-effects model (instead of a fixed-effect model). The results (see right column of Figure S1 in Appendices) showed that the inflation of Type-1 error rates with the flexible stopping rule is more remarkable than for the previous simulations with fixed-effect model. With the random-effects model, the average of the true effect sizes across the studies approaches zero, but the true effect size for each study deviates from zero. The flexible stopping rule seems to take advantage of this between-studies fluctuation of true effect size. It should be noted that, under random effects model, we need to be careful to interpret the statistical test of overall (or integrated) effect size, as this model assumes that the true effect sizes are different (heterogeneous) between studies. In other words, the overall effect size may not represent the true effect size of any of the studies included in the meta-analysis (Borenstein et al., 2009). Nevertheless, in practice, especially in the context of internal meta-analysis, it is common that researchers draw an overall conclusion from the integrated effect size, even if between-studies variance is statistically significant. Finally, the increase in Type-1 error rates were even larger when a researcher used a fixed-effect model meta-analysis to integrate the studies (i.e., when a researcher used a misspecified model) than when using a random-effects model meta-analysis.

Simulation 2: When the first study was a false-positive

It is common practice to sometimes make replication attempts when a first study is statistically significant, in order to ensure that the observed effect is not a false-positive. This attempt is a good practice in itself. However, if the second study is unfortunately not statistically significant but it shows a similar trend to the first study, researchers may be tempted to conduct an internal meta-analysis to examine whether the combined results would yield a statistically significant effect. Does such a use of internal meta-analysis help to eliminate false-positive findings? To address this point, we flagged cases where the 1st study exhibited a (falsely) statistically significant result (this should happen 5% of the total repetitions). For each condition, we ran simulations until we obtained 15,000 (falsely) significant results (i.e., 15,000 Type-1 errors), and then tracked how many of these 15,000 replications would falsely survived statistical significance in combination with the further follow-up studies. The data generation procedure followed a fixed-effect model. We also manipulated the number of participants for each study (N = 20, 30, or 50). The simulation results (Figure 2) showed that more than 40% of these initially statistically significant (by chance) cases survived statistical tests with internal meta-analysis (note that this y-axis is not a Type-1 error rate, as the simulation focuses on the cases where the first study happened to be statistically significant) if a researcher used the flexible stopping rule (left column of Figure 2). Interestingly, even without the flexible stopping rule (right column of Figure 2), about 22-50% of the initial false-positive cases stayed significant after the 2nd study. This rate decreases with additional studies, but even after 8th study, the rate is slightly higher than 5%. This happens because a replication attempt was motivated only *after* seeing the significant result in the first study, which could cause an implicit publication bias. Therefore, the use of internal meta-analysis in such a scenario can certainly reduce the relative risk of false positives, decreasing the overall possibility of false-positive findings (again, note that the y-axis

in Figure 2 is not a Type-1 error rate). However, the effectiveness of internal meta-analysis to eliminate potential false-positive findings is not as strong as researchers would expect (e.g, 5%), especially when they use the flexible stopping rule. When a researcher comes across a false-positive result in the first study, and then runs a replication study, it is somewhat difficult to arrive at the correct conclusion (i.e., non-significant effect), if the researcher relies only on internal meta-analysis. Often one or two additional studies and the use of internal meta-analysis to integrate the results are not sufficient to override the Type I error.



a

Figure 2. Likelihood of obtaining a false-positive result when the flexible stopping rule is used (i.e., researchers stop collecting data upon obtaining significance, $p < .05$) in the left half, conditioned to the cases that showed significant results in the 1st study. The right panels show the likelihood of a false-positive result without the flexible stopping rule (i.e., researchers run studies until the intended number of studies are completed irrespective of statistical significance). The analysis methods are written in the parentheses.

Literature review

Sampling

Now that we have empirically demonstrated an inflation of false positive rates from the flexible stopping rule in an internal meta-analysis, it should be informative to review existing empirical studies to gauge the awareness and possible prevalence of this issue in psychology. We limited our literature review to the papers that cited G. Cumming's seminal work, '*The new statistics: Why and how*' (Cumming, 2013). We used Google Scholar, and retrieved 192 articles which cited Cumming (2013; retrieved at 24th March, 2015). The first author read 188 of these (four articles were written in Spanish and we could not read them), and selected the 31 papers that conducted a meta-analysis. Out of these 31 papers, 16 papers used an internal meta-analysis whilst 15 papers used a meta-analysis of past studies (including their own studies – i.e., cumulative meta-analysis). We focused only on the 16 papers that utilized an internal meta-analysis.

Coding

The 16 papers that utilized an internal meta-analysis were coded in terms of the following aspects, which are presented in Tables 1 and 2. Table 1 presents the information regarding: (1) whether a fixed-effect model or a random-effects model meta-analysis was chosen; (2) the number of studies included in the internal meta-analysis, and (3) whether justifications were made for the number of reported studies. Table 2 mainly presents the results of the statistical tests with the individual studies and internal meta-analysis, including: (4) actual *p*-values of individual studies; (5) the *p*-value of the integrated effect in internal meta-analysis; (6) consistency across the meta-analyzed studies and internal meta-analysis in terms of the significance of the NHSTs; (7) an heterogeneity measure *Q* and Higgins *I*-squared (Higgins, Thompson, Deeks, & Altman, 2003) in terms of the effect sizes across individual studies; (8) and

the post-hoc power based on the integrated effect size and the total sample size. Not all papers reported meta-analysis statistics (e.g., *Q* and *I*-squared); for these papers, we conducted meta-analysis based on the available information and obtained information.

Table 1. Summary of the psychological articles with an internal meta-analysis (citing Cumming, 2013) in terms of the models being chosen and the justification of the choice & the number of studies within a paper

Paper ID	Model	Number of studies in meta-analysis	Justification of the number of studies to conduct
1	Random-effects	2	None
2	Not mentioned	2	None
3	Not mentioned	2	None
4	Not mentioned	2	None
5	Fixed-effect	2	None
6	Random-effects	3	None
7	Raw data pooled	3	None
8	Not mentioned	4	None
9	Not mentioned	4	None
10	Random-effects	3	None
11	Not mentioned	3	None
12	Random-effects	2	None
13	Not mentioned	3	None
14	Not mentioned	2	None
15	Not mentioned	5	None
16	Random-effects	9	None

Note. In Paper ID 7, ANOVA on the pooled data across 3 studies was conducted without accounting for the clustering of participants within studies.

Table 2. Consistencies across individual studies and internal meta-analysis in terms of statistical significance, p-values, and effect sizes: Summary of the psychological articles with an internal meta-analysis (citing Cumming, 2013)

Paper ID	*p*-values of the effects in individual studies					*p*-value of the effect in internal meta-analysis	All the studies and internal meta-analysis were consistent in terms of significance?	*Q*-index (d.f.)	*I-squared*	Power
	1st study	2nd study	3rd study	4th study	5th study					
1	.099+	< .001+	-	-	-	.014 *	No	0.25 (1)	0%	.69
2	.114 .056 +	.005*	-	-	-	< .001 * #	No	- #		- #
3	.320	.07 +	-	-	-	.058 +	No	0.56 (1)	0%	.47
4	.121	.194	-	-	-	.04 *	No	< 0.01 (1)	0%	.54
5	.259	.137	-	-	-	.07 +	No	< 0.01 (1)	0%	.48
6	around .055+ a	around .80 a (opposite direction)	around .20 a	-	-	around .055 +	No	around 2.21 (2) a	9.5% a	around .55 a
7	< .001*	.12	.32	-	-	.001 *	No	-	-	< .78b
8	.097 +	.084 + .32 .002*	.001* < .001*	.31 < .001*	-	< .001 * #	No	- #	- #	> .99 #
9	< .001*	< .001*	< .001*	< .001*	-	<. 001 *	Yes	- #	- #	> .99
10	< .001*	.002*	< .001*	-	-	< .001 *	Yes	0.06 (2)	0%	> .99
11	< .001*	< .001*	< .02*	-	-	< .01 *	Yes	- #		> .99
12	.001*	.003*	-	-	-	< .001 *	Yes	0.0	0%	.98
13	< .001*	< .001*	.043*	-	-	< .001 *	Yes	5.42 (2) +	63.0%	> .99
14	< .001*	.004*	-	-	-	< .001 *	Yes	0.67 (1)	0%	> .99
15	< .001* < .001* < .001* < .001*	< .001* < .001*	.02	.001 .02	< .001*	< .001 *	Yes	- #	- #	> .99
16	9 Studies: *p*s =.700, .718, .830, .883, .159, .05~.06 +, .644, .925, .095					.29 (*n.s.*)	Almost yes	6.983 (8)	0%	> .99

Note. + $p < .10$, * $p < 0.05$. For some studies, the table has more than one *p*-value for each column. This means that the hypothesis was tested with multiple measurements (e.g., different questionnaires).

In these papers, the effects within a study were transformed before submitted to internal meta-analysis (e.g., averaged from multiple dependent variables) but the transformed score and its variance were not reported. Therefore it was impossible to know the exact *p*-values/*Q*-index/power.

a In Paper ID 6, these *p*-values/*Q*-index/power were calculated on the basis of the means & 95%CIs that we measured from the figure (using a ruler). The means & 95%CIs in the table did not allow us to track their analyses as neither a correlation nor a paired *t*-value (and its associated *p*-value) was reported.

b In Paper ID 7, the exact power calculation was impossible due to missing information.

Outcomes

Among these 16 papers, half of them obtained the consistent pattern of findings in terms of the significance of the effect across individual studies and in the final internal meta-analysis (lower half of Table 2). Given the consistency of the individual studies, the internal meta-analysis in these papers can be deemed as a method to provide a narrower confidence interval, rather than a method to make a dichotomous judgment on statistical significance, in line with Cumming's (2008, 2012, 2013) recommendations (see also the results of power analyses, below).

In contrast, for the remaining half (eight) of the papers, the reported results were mostly inconsistent across the studies and internal meta-analysis (upper half of Table 2). One noteworthy observation is that none of these studies provided justifications for the number of studies conducted or declared that they determined the number of studies a priori, indicating little awareness of the issue of the flexible stopping rule in their internal meta-analyses. In addition, all of these studies claimed the statistical significance or non-significance of the overall effect based on the internal meta-analysis results. Among these eight papers, six papers (Paper IDs 1-5, and 8) observed a non-significant effect in the first place, but an additional second study (or more studies in Paper ID 8) made the results from an internal meta-analysis significant (*p*s = below/around .05). These papers did not report further studies. Another paper (Paper ID 6) did not observe a significant effect across three studies but the internal meta-analysis showed a statistically significant effect. If the termination of the studies had been motivated by a significant effect in internal meta-analysis, Type-1 error rates for these studies would be more than the nominal value (i.e., 5%), as indicated in our Monte-Carlo simulation. For the other paper (Paper ID 7), their first study obtained a significant effect but they terminated data collection after obtaining non-significant effects in later studies. Although our simulations did not directly

examine these situations (i.e., we focused on situations where researchers stop collecting data once (pooled) results become statistically significant), if researchers had been motivated to stop collecting data given the overall significant effect from internal meta-analysis, this should also inflate possible Type-1 error rates.

We further examined these papers by performing a power analysis and by checking whether internal meta-analysis substantially reduced *p*-values or not. This approach is inspired by TSA, which has been developed in clinical trials (Brok et al., 2008; Wetterslev et al., 2008). As mentioned in the Introduction, the basic idea of TSA is that we need to consider the statistical power of meta-analysis in order to decide whether it is appropriate to stop running studies. For each paper of Table 2, post-hoc statistical power was calculated on the basis of the identified effect size and the total number of participants in internal meta-analysis (alpha = 0.05). The analysis (rightmost column of Table 2) showed that the statistical power was relatively low for almost all of the papers that showed inconsistent results (IDs 1-7). Although we did not formally use TSA, as not all the papers reported sufficient information to run TSA, these results suggest that, had they used flexible stopping rules, the authors had stopped running studies before sufficient evidence was accumulated to make a valid decision. To elaborate this point, Table 2 also presents the *p*-values of the individual studies and that of internal meta-analysis (some *p*-values were not reported in the original papers, but we attempted to reproduce *p*-values as precisely as possible from the available information --- see footnote of Table 2). The rationale behind this TSA approach is that if statistical power is low, then claiming a statistically significant effect in meta-analysis requires a more strict test-statistic than the conventional nominal level (i.e., *p*-value lower than 5%). In other words, if the final computed *p*-value (unadjusted) is just below or around 5%, researchers may need more studies to draw a solid

conclusion about the effect (i.e., the current result may be spurious). Our results indicate that, for the papers that showed inconsistent results, the final *p*-value computed from internal meta-analysis was indeed in many cases just below (or around) 5% (e.g., Paper IDs 1, 3-7, and possibly 2), not being substantially reduced by internal meta-analysis. In contrast, the papers that showed consistent results (IDs 9-16) had a strong statistical power in their internal meta-analysis, supporting their decision to quit conducting an additional study.

We must emphasize that we by no means claim that any of these papers actually used the flexible stopping rule (be it implicitly or explicitly) to obtain statistically significant results --- this is simply indeterminable based on the information provided by the papers. Further, if researchers did not use flexible stopping rules (or other questionable research practices), insufficient statistical power alone does not lead to the inflation of Type-1 error rates (Murayama, Pekrun, & Fiedler, 2013). Nevertheless, our literature review and statistical power analyses suggest that we cannot eliminate the possibility that internal meta-analysis was used with the motivation to achieve statistically significant results.

We also found that nine of the 16 papers did not report whether they chose a fixed-effect or a random-effects model meta-analysis. Some of these studies were likely to have chosen a fixed-effect model, as they reported statistical results which could be used for a justification of adopting a fixed-effect model (e.g., a study indicated the heterogeneity index *Q* was non-significant; see Borenstein et al., 2009, for the problem to rely on the *Q* index) but the selection of meta-analysis model was not explicitly specified. Overall, *Q* index did not suggest substantial heterogeneity across studies, but quite a few studies did not report *Q* index or did not have sufficient information to compute *Q* index. As indicated above, we also noted that in some of the cases (seven out of 16 papers) the exact statistics being submitted to the meta-analysis were not

explicitly reported (see footnote of Table 2). In other cases (four out of the 16 papers) the effect sizes in each study and their CIs were drawn in a figure, but the exact values were not reported.

Discussion

Since Cumming (2008, 2012, 2013), internal meta-analysis has attracted increasing attention and seemed to gain the status of a recommended reporting strategy of research findings. We believe this trend will be further accelerated, in light of the fact that recent large-scale reproducibility projects (Klein et al., 2014; Open Science Collaboration, 2015) have shown the utility of meta-analysis to demonstrate replicability in psychology. With this background, the current manuscript provided a cautionary note about the use of internal meta-analysis, pointing out the possible inflation of false-positive rates if researchers are motivated to obtain statistical significance from the meta-analysis, and discussed the issue both using statistical simulations and by conducting a literature review. Our simulation revealed that the flexible stopping rule indeed inflated false positive error rates (Figures 1 and 2). Moreover, fixed-effect model meta-analysis was generally more susceptible to inflation of Type-1 error rates when the flexible stopping rule was employed. From the review of the empirical papers that employed an internal meta-analysis, we found that eight of the 16 papers seemed to use their internal meta-analysis with the aim to deliver a more precise estimation (i.e., narrower CI) for the population parameter, as Cumming emphasized (Cumming, 2008, 2012, 2013). In fact, they established a strong power at the point when they stopped further studies. In contrast, the remaining eight papers seemed to perform internal meta-analyses as a tool to draw a dichotomous decision, and most of these papers did not achieve sufficient statistical power in their internal meta-analysis. As none of the papers described the rationale to determine the number of conducted studies, we could not

exclude the possibility that the researchers might have acted (perhaps inadvertently) in accordance with the flexible stopping rule to obtain a significant effect in internal meta-analysis.

Possible strategies to control type-1 error inflations in internal meta-analysis

Again, we need to emphasize that the issue being discussed here does not lie in (internal) meta-analysis per se, but in researchers' (implicit) motivation to use the strategic flexible stopping rule to maximize positive results with NHST (see also, Ioannidis, 2005; Klein et al., 2014; Open Science Collaboration, 2012, 2013, 2015; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012). Without using the flexible stopping rule, one can reliably keep false positive error rates to the conventional nominal level (e.g., 5%) regardless of how many studies are integrated within a paper. Importantly, Cumming's (2008, 2012, 2013) 'new' statistics and its accompanying recommendation for internal meta-analysis actually emphasized the value of estimation (i.e., CI), and he never recommended making a dichotomous decision based on NHST with internal meta-analyses (see also, Trafimow & Marks, 2015).

One important message from this paper is the importance of transparency of how researchers decide when to stop conducting further replications. As long as the decision to stop is not based on the data themselves, we do not need to worry about the issues discussed in this paper. Even if the decision is based on the previous results, transparency helps readers/reviewers gauge the extent of the seriousness of the issue (Murayama et al., 2013; Wigboldus & Dotsch, 2015). For example, psychologists can consider statistical methods to control Type-1 error inflation in cumulative meta-analysis, such as a power analysis and/or adjustment of the *p*-value (or alpha) using statistical techniques (TSA: Brok et al., 2008; Wetterslev et al., 2008; for extensions of TSA, see, Miladinovic et al., 2013; Thorlund et al., 2009; semi-Bayesian approach: Higgins et al., 2011; sequential meta-analysis with triangular boundaries: van der Tweel &

Bollen, 2010; recalculation of sample size after each cumulative meta-analysis: Roloff et al., 2013). We also hope that our empirical demonstration of the inflation of false positive error rates will be helpful for researchers to evaluate the extent to which the nominal alpha level should be adjusted.

Another simpler and perhaps easier solution would be to make judgments on statistical significance based solely on the (high-powered) replication studies, and make use of the internal meta-analysis only as a way to estimate the CI of effect size. This way, statistical tests are not influenced by the results of previous studies within a paper while preserving the spirit of internal meta-analysis as recommended by Cumming (2012, 2013). This is particularly advisable when the replication study was motivated after obtaining a significant effect in the 1st study. As our statistical simulation (Simulation 2) indicated, when the first study happens to be a false-positive, internal meta-analysis may not be powerful enough to override it with an additional study. These findings also have some implications about the recent large-scale replication study (Open Science Collaboration, 2015). Specifically, this replication study conducted a cumulative meta-analysis to integrate the effect from the original study and that from the pre-registered replication study, and found that the effect size of the latter was much smaller than that of the former in 82.8% of the articles. Nonetheless, combining original and replication results left 68% with statistically significant effects. The authors speculated that some forms of biases in the original published studies may explain the relatively high rate of significant results (see also, Nuijten, van Assen, Veldkamp, & Wicherts, 2015, for a similar warning message about meta-analyzing the biased data from published papers), but these results may also indicate the limitation of conducting a meta-analysis contingent upon the significant results of the first study to perfectly

eliminate potential false-positive studies, although it can certainly reduce the overall false-positive rates.

It is also advisable that the detailed statistics in individual studies (e.g., *p*-values, effect size and its variance) that are submitted to internal meta-analysis should be described sufficiently enough that readers can replicate the meta-analysis. We observed some studies where the way the data were analyzed for each study was different from the way effect sizes were computed and integrated in the internal meta-analysis (e.g., a study had three depending variables and conducted a meta-analysis with the averaged scores of these three variables, without reporting the variance of the averaged variable). In relation to this, a number of studies did not specify the model choice (fixed-effect model vs. random-effects model meta-analyses) or did not provide justifications for their choice of model. These different ways of conducting meta-analyses may also provide researchers with leeway to exploit analysis to attain statistically significant results (a version of "researchers' degrees of freedom"; see Simmons et al., 2011), making internal meta-analyses more susceptible to the inflation of Type-1 error rates.

The use of Bayes factor may be another viable option (Higgins et al., 2011). This approach can tell the relative plausibility of a hypothesis in comparison with an alternative hypothesis, which is independent of a sampling plan (Andraszewicz et al., 2015; Rouder, 2014). Specifically, Bayes factor quantifies the relative strength of the extent to which one's belief (hypothesis) has been supported by the data. Importantly, the rationale of this approach is to continue sampling the data, and when the relative strength of either hypothesis becomes large enough (larger than a conventional critical value), the researcher stops collecting data (Andraszewicz et al., 2015). This means that this approach does not require a pre-determined sample size. Rather, the approach encourages data-monitoring during collection (Rouder, 2014).

Indeed, Andraszewicz et al. (2014) generated three versions of randomized orders in which their (real) data were collected, and demonstrated that the final conclusion was the same as far as the data-collection was continued until a compelling Bayes factor was reached.

Another possible solution would be pre-registering (Chambers, 2013) multiple studies in advance. Preregistration is one of the most effective ways to ensure research transparency. Although pre-registration is typically associated with single high-powered studies, it is possible to pre-register multiple studies that slightly change the key manipulation/population to test robustness of the findings. Importantly, this approach does not require high statistical power for each study --- the power calculation should be made based on the planned studies as a whole (for power calculation in meta-analysis, see Hedges & Pigott, 2001). Therefore, this approach is not as costly as it sounds (because the required total sample size would not be much different from a single high-powered study) while having a handsome benefit of increasing the generalizability of the tested hypothesis.

Conclusion

Over the last two decades it has become increasingly common to include multiple studies in a single paper (Sherman, Buddie, Dragan, Christian, & Finney, 1999). There may be many reasons to include multiple studies in a single paper, but one common reason is to establish the robustness and replicability of the findings. The importance of replication has recently been further highlighted by the recent “replication crisis” in psychology, and researchers in the future will likely be required by journal editors or reviewers to include replication studies in a single paper. With multiple studies in a single paper, conducting an internal meta-analysis seems to be a reasonable choice to make a scientifically robust overall conclusion of the study. This is indeed true --- internal meta-analysis is a good tool to provide the most accurate estimates of effects and

narrowest confidence intervals. It can be a tool for good science. However, when used inappropriately (i.e., used to decide whether continue or discontinue the research based on statistical significance), it is not difficult to draw a false conclusion (i.e., false-positives) that the researcher wanted to support. This can happen even when researchers are not aware of the problem of flexible stopping rule, and have used internal meta-analysis with an honest intent of good scientific practice. We are hoping that our paper will give good guidance for future studies that intend to use internal meta-analysis.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., . . . Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556-578. doi:10.1177/1745691614545653

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An Introduction to Bayesian Hypothesis Testing for Management Research. *Journal of Management, 41*(2), 521-543. doi:10.1177/0149206314560412

Berkey, C. S., Mosteller, F., Lau, J., & Antman, E. M. (1996). Uncertainty of the time of first significance in random effects cumulative meta-analysis. *Controlled Clinical Trials, 17*(5), 357-371. doi:10.1016/s0197-2456(96)00014-1

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis (Statistics in Practice)* Chichester: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111. doi:10.1002/jrsm.12

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science, 9*(3), 333-342. doi:10.1177/1745691614529796

Brok, J., Thorlund, K., Gluud, C., & Wetterslev, J. (2008). Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology, 61*(8), 763-769. doi:10.1016/j.jclinepi.2007.10.007

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci, 14*(5), 365-376. doi:10.1038/nrn3475

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. doi:http://dx.doi.org/10.1016/j.cortex.2012.12.016

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243-253.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychol Methods, 14*(2), 165-176. doi:10.1037/a0015565

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*(4), 286-300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, England: Routledge.

Cumming, G. (2013). The New Statistics: Why and How. *Psychological Science*. doi:10.1177/0956797613504966

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177-188. doi:http://dx.doi.org/10.1016/0197-2456(86)90046-2

Donnellan, M. B., Lucas, R. E., & Cesario, J. (2015). On the Association Between Loneliness and Bathing Habits: Nine Replications of Bargh and Shalev (2012) Study 1. *Emotion, 15*(1), 109-119. doi:10.1037/a0036079

Egger, E., ZellwegerZahner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *Lancet, 350*(9074), 326-329. Retrieved from <Go to ISI>://WOS:A1997XP24200010

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher, 5*(10), 3-8. doi:10.3102/0013189x005010003

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 203-217.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ : British Medical Journal, 327*(7414), 557-560. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC192859/

Higgins, J. P. T., Whitehead, A., & Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in Medicine, 30*(9), 903-921. doi:10.1002/sim.4088

IntHout, J, Ioannidis, J. P, & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*. 14:25. doi:10.1186/1471-2288-14-25.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*. doi:10.1371/journal.pmed.0020124

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152. doi:http://dx.doi.org/10.1027/1864-9335/a000178

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 79*, 659-663.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction. *New England Journal of Medicine, 327*(4), 248-254. doi:doi:10.1056/NEJM199207233270406

LeBel, E. P. (2015). A new replication norm for psychology. *Collabra*, 1(1), doi:http://dx.doi.org/10.1525/collabra.23

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics, 55*(4), 1286-1290. doi:10.1111/j.0006-341X.1999.01286.x

Müller, H. H., & Schafer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics, 57*(3), 886-891. doi:10.1111/j.0006-341X.2001.00886.x

Miladinovic, B., Mhaskar, R., Hozo, I., Kumar, A., Mahony, H., & Djulbegovic, B. (2013). Optimal information size in trial sequential analysis of time-to-event outcomes reveals potentially inconclusive results because of the risk of random error. *Journal of Clinical Epidemiology, 66*(6), 654-659. doi:10.1016/j.jclinepi.2012.11.007

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why Hypothesis Tests Are Essential for Psychological Science: A Comment on Cumming (2014). *Psychological Science, 25*(6), 1289-1290. doi:10.1177/0956797614525969

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research Practices That Can Prevent an Inflation of False-Positive Rates. *Personality and Social Psychology Review, 18*(2), 107-118. doi:10.1177/1088868313496330

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19*(2), 172-182. doi:http://dx.doi.org/10.1037/gpr0000034

O'Brien, P. C., & Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics, 35*(3), 549-556. doi:10.2307/2530245

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

Open Science Collaboration. (2013). The reproducibility project: A model fo large-scale colalboration for empirica research on reproducibility *Implementing reproducible computational research (a volume in the r series)*. New York, NY: Taylor & Francis.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. . *Perspectives on Psychological Science, 7*(6), 531-536

Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science, 7*(6), 528-530. doi:10.1177/1745691612465253

Pocock, S. J. (1977). Group sequential methods in the design and analysis of

clinical trials. *Biometrika, 64*, 191-199.

Pogue, J., & Yusuf, S. (1997). Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control Clin Trials, 18*, 580-593.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal, 340*. doi:10.1136/bmj.c221

Roloff, V., Higgins, J. P. T., & Sutton, A. J. (2013). Planning future studies based on the conditional power of a meta-analysis. *Statistics in Medicine, 32*(1), 11-24. doi:10.1002/sim.5524

Rosenthal, R. (1979). The file drawer problem and roleance for null results. *Psychological Bulletin, 86*(3), 638-641. doi: http://dx.doi.org/10.1037/0033-2909.86.3.638

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301-308. doi:10.3758/s13423-014-0595-4

Schwarzer, G. (2012). meta: meta-analysis with R (Version Version R package Version 2.1–1.). Retrieved from http://CRAN.R-project.org/package=meta

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin, 105*(2), 309-316. doi:10.1037//0033-2909.105.2.309

Sherman, R. C., Buddie, A. M., Dragan, K. L., Christian, M. E., & Finney, L. J. (1999). Twenty years of PSPB: Trends in content, design, and analysis. *Personality and Social Psychology Bulletin, 25*(2), 177-187.

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*(11), 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. doi:http://dx.doi.org/10.1037/a0033242

Thorlund, K., Devereaux, P. J., Wetterslev, J., Guyatt, G., Ioannidis, J. P. A., Thabane, L., . . . Gluud, C. (2009). Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International Journal of Epidemiology, 38*(1), 276-286. doi:10.1093/ije/dyn179

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. doi:10.1080/01973533.2015.1012991

van der Tweel, I., & Bollen, C. (2010). Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials, 7*(2), 136-146. doi:10.1177/1740774509360994

Wetterslev, J., Thorlund, K., Brok, J., & Gluud, C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology, 61*(1), 64-75. doi:10.1016/j.jclinepi.2007.03.013

Wigboldus, D. J., & Dotsch, R. (2015). Encourage Playing with Data and Discourage Questionable Reporting Practices. *Psychometrika*, 1-6. doi:10.1007/s11336-015-9445-1