Exploring Small, Confirming Big: An Alternative System to *The New Statistics* for Advancing Cumulative and Replicable Psychological Research

John Kitchener Sakaluk

Department of Psychology

University of Toronto Mississauga

I am sincerely grateful for the support and feedback provided by Monica Biernat on earlier drafts of the manuscript, and the other faculty and graduate students of the Social Psychology program at the University of Kansas for their feedback on the concept for the paper.

Correspondence concerning this article should be addressed to John Sakaluk, 3359 Mississauga Road, Deerfield Hall Rm. 4098, Mississauga, Ontario, Canada L5L IC6

Email: john.sakaluk@utoronto.ca

Abstract

While outlining his vision of *The New Statistics*, Cumming (2014) proposes that a more rigorous and cumulative psychological science will be built, in part, by having psychologists abandon traditional null-hypothesis significance testing (NHST) approaches, and conducting small-scale meta-analyses on their data whenever possible. In the present paper, I propose an alternative system for conducting rigorous and replicable psychological investigations, which I describe as *Exploring Small, Confirming Big*. I begin with a critical evaluation of the merits of NHST and small-scale meta-analyses, and argue that NHST does have a valuable role in the scientific process, whereas small-scale meta-analyses will do little to advance a cumulative science. I then present an overview of an alternative system for producing cumulative and replicable psychological research: *Exploring Small, Confirming Big*. It involves a two-step process to psychological research, consisting of: (1) small *N* investigation(s), in which psychologists use NHST to develop exploratory models; and (2) strong, confirmatory tests of exploratory models, by analyzing new and/or existing large *N* datasets with variables that capture the effect(s) of interest from the *Exploring Small* stage. I conclude by discussing several anticipated benefits and challenges of adopting the *Exploring Small, Confirming Big* approach.

Keywords: big data; data analysis; meta-analysis; open science; replicability

Exploring Small, Confirming Big: An Alternative System to *The New Statistics* for Advancing Cumulative and Replicable Psychological Research

Recent investigations into the replicability of psychological science, as well as high profile unveilings of fraud, have placed our discipline—and many others—in a state of somber introspection. The issues of fraud (Crocker & Cooper, 2011; Simonsohn, 2013), misreporting (Bakker & Wicherts, 2011; Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007; Wicherts, Bakker, & Molenaar, 2011), p-hacking and other questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), the dearth of replication (Makel, Plucker, & Hegarty, 2012), and other possible mechanisms that might render a published finding "false" (cf. Ionnidis, 2005), are on the minds of many psychologists.

Psychologists have subsequently proposed a number of potential solutions to metascientific concerns, in an attempt to reassure themselves—and the broader public (e.g., Carey, 2011)—of the value and replicability of our science. Many, for example, are calling on colleagues to make their data more accessible (e.g., Crocker & Cooper, 2011; Simonsohn, 2013). Others proposals include publishing fewer papers (Nelson, Simmons, & Simonsohn, 2012), so that research might be carried out more carefully; large-scale collaborative replication efforts (e.g., Klein et al., 2014), to rigorously test some of the foundational findings of our field; pre-registration of hypotheses and data analysis plans (Wagenmakers, Wetzels, van der Maas, & Kievit, 2012), to keep psychologists honest against their motivated reasoning that is natural throughout the research process; making adjustments to the standard peer-review process (Sakaluk, Williams, & Biernat, 2014), in order to catch reporting errors that can, and often do happen by mistake; or, as a seemingly last-ditch solution, using statistical methods to detect and steer clear of effects that might be particularly unlikely to replicate (Simonsohn, Nelson, &

Simmons, 2014). Each of these proposals to increase replicability (and many others) is unique in their merits and challenges for implementation.

Cumming's proposal for psychologists to adopt *The New Statistics* (TNS; 2014), however, stands alone as the most, far-reaching, and widely discussed solutions to the replicability concerns within psychology. In his landmark article, Cumming makes twenty-five recommendations for improving the way that psychologists conduct research (Cumming, 2014); many of these, I think, are excellent suggestions. Even so, Cumming makes two recommendations that, in my opinion, are not conducive to improving psychological science. Perhaps somewhat predictably, they are also the recommendations that demand the greatest amount of disciplinary reform. They include: "10. Whenever possible, avoid using statistical significance or *p* values; simply omit any mention of null-hypothesis significance testing (NHST)" and "18. Use small- or large-scale meta-analysis whenever that helps build a cumulative discipline." (p. 2)

The recommendations of Cumming (2014) that I believe to be problematic would not be so troublesome, from my perspective, if they simply existed at the stage of methodological proposal. Journals in our field, however, are already beginning to adopt TNS, without there having been much opportunity to debate the specific merits and limitations of TNS as a system of research.

The purpose of this paper is therefore twofold. First, after summarizing Cumming's (2014) positions against NHST and in favor of small-scale meta-analysis, I offer some arguments in favor of keeping NHST as a part of psychologists' statistical repertoire, and against the widespread adoption of small-scale meta-analysis. I then propose a new approach for advancing replicable research in psychology, as an alternative to TNS. The proposed system, which I describe as *Exploring Small, Confirming Big*, allows researchers to use exploratory NHST practices for initial studies involving relatively smaller sample sizes; researchers then perform strong confirmatory tests of their exploratory findings using more restrictive, large sample size datasets. In the second portion of the paper, I therefore lay out the general *Exploring Small, Confirming Big* approach. I conclude by addressing some anticipated benefits and challenges of implementing the *Exploring Small, Confirming Big* approach.

On NHST in *The New Statistics*

Of the 25 proposed changes made in TNS, recommending that psychologists cease using NHST is easily the largest in scope, and therefore the one that psychologists are likely to be resisting the most. Nevertheless, one major social psychology-related periodical—*Psychological Science*—now discourages the use of NHST (Eich, 2014), and another—*Basic and Applied Social Psychology*—has banned NHST (Trafimow & Marks, 2015). I therefore first begin with a review of Cumming's (2014) arguments against NHST, while paying particular to his arguments against defenses of NHST. I follow by then highlighting at least three instances in which NHST does appear to advance a cumulative and replicable psychological science.

No Room for NHST in *The New Statistics*

Cumming (2014) made abundantly clear that he saw no place for NHST in psychological research. He drew on critiques of NHST offered by the likes of Fidler (2005), Kirk (2003), and Kline (2004), who highlighted the many conceptual, logical, and practical shortcomings of NHST. I have no interest in contending with any of these criticisms, they are largely, if not entirely, correct. NHST is a statistical paradigm fraught with limitations. What I do disagree with, however, is the recommendation that, "The best policy is, whenever possible, not to use NHST at all." (Cumming, 2014, p. 12)

Cumming (2014) gave somewhat short shrift to defenses of NHST, essentially punting the matter to a chapter written by Schmidt & Hunter (1997), in which the authors examined eight common objections to the discontinuation of NHST. As with the papers critiquing NHST, Schmidt and Hunter's (1997) rebuttals to most of these objections are thorough and convincing. Yet there are a few aspects of this chapter—and therefore the foundation of TNS—that I find problematic.

First, consider the following summary of NHST offered by Schmidt and Hunter (1997): "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution." (p. 3, emphasis added) Theirs is plainly a black and white characterization of the use of NHST. But isn't this precisely the type of dichotomous scientific thinking that *both* Cumming (2014) and Schmidt and Hunter (1997) encouraged psychologists to avoid? I submit that the use of NHST likely ranges in its utility depending on the analytic circumstances, and it may be more important to assess under what circumstances NHST can be useful, and I will soon advocate for particular uses of NHST that benefit scientific progress.

There is also an important contradiction in Schmidt and Hunter's (1997) rebuttals to two of the defenses of NHST that they reviewed: "…the problem is not use but misuse of significance testing," (p. 20) and, "…the effort to reform data analysis methods should be dropped because it is futile." (p. 22). The authors first argued that it would be impossible to convince psychologists to use and interpret NHST correctly, because these mistaken practices are too deeply engrained in conventional practice. They later followed by deploring cynical attitudes about the possibility of convincing psychologists to abandon the use of NHST altogether suggesting that such a sense of hopelessness is "corrosive of scientific values" (p. 26). It is unclear to me as to why statistical mavens (Sharpe, 2013) should concede the battle of correct usage of NHST, while continuing to fight for the abandonment of NHST—a much loftier goal. I remain optimistic on both fronts: with greater attention on the use and misuse of NHST, hopefully we will see gradual improvement in the interpretation of NHST-based findings, as well as greater uptake of non-NHST statistical paradigms that are better suited to addressing the research questions of psychologists (e.g., Bayesian, Kruschke, 2014).

To be clear, I wholly agree with many of the established limitations of NHST, highlighted by those cited in Cumming's proposal of TNS (2014). I am similarly convinced of the speciousness of many—but not all—of the defenses of NHST, discussed by Schmidt and Hunter (1997). Yet I think to advance that "significance testing never makes a useful contribution to the development of cumulative knowledge," (p. Schmidt & Hunter, 1997, p. 22) goes too far. Indeed such an absolute position is more vulnerable to counterargument; I now turn to highlighting three instances in which NHST does, in fact, help to build a cumulative and replicable psychological science.

Ways that NHST *Does* Help to Build a Cumulative Science of Psychology

In fairness to Cumming (2014), he later qualifies his recommendation to abandon NHST with the following:

> I include "whenever possible" in my recommendations that we avoid NHST, to cover any cases in which it is not possible to calculate a relevant CI; I expect such cases to be rare, and to become rarer. (p. 26)

Unfortunately, I think such instances are more common than Cumming estimates, thus greatly reducing the helpfulness of his recommendation to abandon NHST. Below are three instances in which NHST helps to build a cumulative and replicable psychological science, and there are likely others I have not considered.

NHST helps psychologists make necessary dichotomous analytical decisions. When studying psychological phenomena, I agree with Cumming's (2014) objections to dichotomous thinking; it is more useful to ask to what extent something exists, and under what conditions, rather than asking the dichotomous question of whether something exists at all. But there are times, both in the course of doing research, and applying research, when a dichotomous decision must be made. Intervene (or not) in the cycle of an individual's substance abuse? Select one model of structural relations between psychological variables over another (or not)? Do, or do not—there is often no interval for initial action.

Interestingly, the need to make dichotomous decisions is one defense of NHST that some authors cited by Cumming (2014) appear to appreciate. Indeed, Kline (2004) discussed the value of NHST in helping to making dichotomous decisions, in his section, "Is There Anything Right with NHST?" (p. 79-82), whereas Fidler (2005) suggested, "such scenarios [dichotomous decision-making] form the more compelling defenses of NHST." (p. 45) Researchers carrying out a meta-analysis, for example, must decide whether and how to correct for the presence of publication bias (Rosenthal, 1979)—a process facilitated by the use of NHST.[1] Likewise, NHST methods represent the dominant approach to selecting between competing nested models (e.g., in the case of structural equation modeling, and multilevel modeling; see Hoyle, 2012, and Snijders & Bosker 1999, respectively). And though NHST alternatives are available for nested model comparisons, they are beset by a number of empirical and practical limitations.[2] These are but two examples of the utility of NHST when making necessary dichotomous decisions that occur throughout the research process; I am certain readers will be able to think of others.

Facilitating p-curve analysis. One additional way that NHST may yet help to build a cumulative and replicable psychological science involves the newly developed *p*-curve technique

(Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2014). Unlike detecting and correcting for publication bias in meta-analysis, or conducting nested models comparisons, *p*-curve isn't helpful because of any need to make dichotomous decisions. Rather, *p*-curve offers new possibilities for how psychologists can use exactly reported *p*-values, in order to assess the evidential value of a given literature (Simonsohn, Nelson, & Simmons, 2014), and to potentially improve on methods of meta-analytic estimation (Simonsohn, Simmons, & Nelson, 2014).

P-curve requires the plotting of the reported exact *p*-values from a given study, article, author, or literature that are lesser than or equal to .05 (Simonsohn, Nelson, & Simmons, 2014). Researchers can then perform a number of tests of skewness on a given *p*-curve, in order to test: (1) whether a set of *p* values contains evidential value; (2) whether a set of *p* values lacks evidential value; and (3) whether a set of *p* values have been intensely *p*-hacked. This ability to assess the lack of evidential value and whether results have been *p*-hacked will likely prove useful when scholars are reviewing literatures, directing replication efforts, or making decisions throughout the peer-review process. Further simulations suggest that *p*-curve may dramatically outperform trim and fill in terms of estimation accuracy of mean effect size when publication bias is present (Simonsohn, Simmons, & Nelson, 2014), though it remains to be seen how *p*-curve compares to more sophisticated methods of meta-analytic estimation (e.g., PET-PEESE, Stanley and Doucouliagos, 2013).

Summary of the utility of NHST. Discussion of NHST within the introduction to TNS (Cumming, 2014, and articles cited therein; e.g., Schmidt & Hunter, 1997) appears to imply that psychological science employing NHST is hardly science at all. I hope I have revealed that this claim does not hold up to closer scrutiny. NHST is indeed limited in a number of important ways (cf. Kline, 2004), but it most certainly still has a role to play in cumulative and replicable psychological science. Namely, it helps researchers to make important dichotomous decisions throughout the process of data analysis. And as the advent of *p*-curve suggests (Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2014), we may find even more creative and helpful uses for the output of NHST. Psychologists should therefore continue to use NHST—at least for the analytic purposes I have discussed—in addition to reporting confidence intervals and effect sizes.[3]

On Small-Scale Meta-analysis in *The New Statistics*

Meta-analyses are a crucial method for achieving a cumulative science of psychology (Chan & Arvey, 2012; Schmidt, 1992). Cumming (2014), I think, therefore does the field a service in stressing the importance of meta-analytic thinking in TNS. However, to suggest that small-scale meta-analysis helps to build a cumulative discipline seems unlikely, at best, and entirely at odds with the goal of building a cumulative and replicable science, at worst.

Arguments Against Small-Scale Meta-Analysis

How small is a small-scale meta-analysis? From Cumming (2014):

> A minimum of two results may suffice for a meta-analysis. Consider meta-analysis to combine results from several of your studies, or from your current study plus even only one or two previous studies." (p. 14)

However, small-scale meta-analyses are not an effective means of achieving the goals of a cumulative and replicable science, for a number of reasons that I now detail.

Garbage in, garbage out. Meta-analyses are only as good as the studies they review, for building a cumulative psychological science. The integrity of small-scale meta-analytic estimates could, for example, be compromised substantively by including studies in which *p*-hacking took place (Simmons et al., 2011). If a psychologists were to *p*-hack three studies to have significant effects, and then conduct a small-scale meta-analysis on these studies, there is a high probability that the resulting meta-analysis will support the *p*-hacked conclusions.[4]

Small-scale meta-analyses are not cumulative. Even when small-scale meta-analyses include a sprinkling of effect sizes from the literature, in addition to effect sizes from the researcher's own studies, small-scale meta-analyses do not contribute to a cumulative and replicable science. Specifically, encouraging researchers to meta-analyze their studies in conjunction with "even only one or two previous studies" (Cumming, 2014, p. 14) will inevitably lead researchers, because of their motivated reasoning or otherwise, to cherry-pick studies in the literature that complement their own results while ignoring ones that don't. Thus, in addition to *p*-hacking, we as a discipline will now need to be concerned with the possibility of *meta-hacking*.

However, small meta-analyses, in which some existing literature was included and some wasn't, will also make it more difficult for large and exhaustive meta-analyses to be conducted. Aspiring large-scale meta-analysts will now have to screen for duplicate effect sizes in multiple small-scale meta-analyses, or ignore small-scale meta-analyses all together. In the case of the former, small-scale meta-analyses actively stall progress towards a cumulative science, whereas in the case of the latter, they simply will not contribute to a cumulative science at all. A worst case scenario may be that the additional burden of sifting through small-scale meta-analyses may discourage researchers from conducting comprehensive meta-analyses, as conducting comprehensive meta-analyses is already logistically onerous (e.g., converting between effect sizes, contacting corresponding authors when insufficient statistical information is reported, searching for unpublished literature, etc.,).

Meta-analyses should be decisive. Somewhat implicit in desiring meta-analyses to be cumulative, is our field's desire for meta-analyses to be decisive. Indeed, chief among the reasons that meta-analyses are often so highly cited is that, when done well, large-scale meta-analyses become the *de facto* source on the current state of knowledge regarding a psychological phenomenon. Numerous small-scale meta-analyses of the same effect will therefore will dramatically undermine the decisiveness-value of meta-analyses. What will happen if small-scale meta-analyses of the same effect produce discrepant findings? Likely, what happened before meta-analyses became commonplace; akin to discrepant findings from individual studies, researchers will be able to selectively pick and choose individual small-scale meta-analyses to cite, based on which best support their theorizing, until/unless someone(s) finally conducts a large-scale and definitive meta-analysis of the phenomenon.

Summary of the utility of small-scale meta-analysis. Small-scale meta-analyses might prove helpful for scholars hoping to convince editors and reviewers their effects from a multi-study paper are "real". I also find Fabrigar and Wegener's appraisal of the utility of small-scale meta-analysis for evaluating the success of replication studies (located in this issue) to be quite promising. Otherwise, however, I see very little benefit in encouraging their use. When conducted *in abstentia* of external literature, small-scale meta-analyses could encourage meta-hacking; when external literature is selectively included, small-scale meta-analyses could encourage biased selection of outside literature, and potentially make large-scale meta-analyses more logistically burdensome to carry out. A cumulative and replicable psychological science would therefore hardly be benefited by the creation of a cottage industry of indecisive, and likely inconsistent, small-scale meta-analyses.

An Overview of the *Exploring Small, Confirming Big* Approach

As an alternative to TNS, I would encourage psychologists to consider following an *Exploring Small, Confirming Big* approach when conducting their research. This two-stage approach to research recognizes that much of the exploration and serendipity of psychological science occurs during the conduct of small *N* studies, while at the same-time, acknowledging that abuse of researcher "degrees of freedom" (Simmons et al., 2011) in small *N* studies demands psychologists to perform stronger confirmatory tests of their exploratory findings (Wagenmakers et al., 2012). Psychologists following an *Exploring Small, Confirming Big* approach would therefore conduct both *Exploring Small* and *Confirming Big* investigations of the same phenomenon, and report on both in the same manuscript in the course of attempted publication.

Exploring Small

What is exploring small? In the *Exploring Small* stage, psychologists go about their scholarly business largely in their usual fashion—conducting small *N* studies and using NHST methods of analysis. Psychologists can also make use of whatever researcher degrees of freedom they wish (e.g., sample size adjustments, exploratory covariates, etc., John et al., 2012; Simmons et al., 2011), as long as they accurately and comprehensively report their data collection and analysis practices. In doing so, psychologists should *always* report confidence intervals and effect sizes for the comparisons and associations they are exploring, as recommended by Cumming (2014) and the APA (2010).

Confirming Big

What is confirming big? In the *Confirming Big* stage, scholars subject the exploratory models they developed in the *Exploring Small* stage, to a strong confirmatory test, through analysis of a large *N* dataset. There may be a number of ways for *Confirming Big*; I now describe some intuitive (i.e., large *N* replications; Many Labs collaborations) and less intuitive (i.e., (inter)national datasets; Google Correlate; analyzing social media data; and meta-*re*analysis) methods for psychologists to use, in order to accomplish this goal.

Intuitive methods for confirming big. Intuitive methods for *Confirming Big* include individually or collectively (e.g., via a Many Labs collaboration) collecting a large N dataset to test the exploratory model determined in the *Exploring Small* stage.[5] Though these conventional methods may lack some of the "flashiness" of the less intuitive methods described later, their primary strength is that they should be applicable for any *Exploring Small* study design—particularly those that would be difficult to confirm big via the less conventional methods.

Individual large N replications. Perhaps the most straightforward way of *Confirming Big* would be for researchers to conduct their own large *N* replications of their exploratory models. In doing so, researchers should preregister their hypotheses, materials, and data analysis plan (e.g., via the Open Science Framework). Further, in order to reduce the temptation to *p*-hack, included measures should consist only of those needed to assess the variables used in the *Exploring Small* model, and those needed to sufficiently describe the characteristics of the *Confirming Big* sample.

Confirming Big samples of this sort could be collected via services like Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011). However, an important consideration this method is determining how large of a sample would be sufficient for being considered a bona fide *Confirming Big* attempt? I am reluctant to impose rigid requirements in this matter, but I think that, at bare minimum, samples for *Confirming Big*, should: (1) possess 80% power (at least) to detect effects of the size found in the *Exploring Small* model, and (2) also appear somewhat superficially large, in a subjective sense.

Consider, for an example, a simple between-subjects experiment consisting of two conditions. In the *Exploring Small* stage, the researcher(s) might rely on minimal conventions for sample size determination, such as the 20-per-cell guideline offered by Simmons and colleagues (2011), recruit 40 participants, and find a standardized mean difference of $d = 0.30$ between the two groups. During the *Confirming Big* stage, the researcher(s) would need to recruit a much larger sample; recruiting 400 participants (200 per condition), for instance, would grant 80% power to detect effects slightly smaller ($d = 0.28$) than those found in the *Exploring Small* stage.

Many labs replications. Should, for any number of reasons (e.g., insufficient funds), an individual large N replication is not a viable strategy for *Confirming Big*, researchers could opt to pursue a "Many Labs" approach to replicating their *Exploring Small* models (see Open Science Collaboration, 2012, Uhlmann et al., this issue, for examples). In adopting this strategy, labs collaborating in the *Confirming Big* process would collectively preregister their hypotheses, materials, and analysis plan. Individual labs would then recruit independent samples that, on their own, would be insufficient for the task of *Confirming Big*, but when combined and analyzed (i.e., meta-analytically), would meet—and potentially exceed—the sample size requirements outlined in the previous section.

Of course, one evident limitation of this approach to *Confirming Big* is that the original researchers must convince some number of their colleagues that their exploratory effect is worth investing the resources to collectively examine. But when successful, this method of *Confirming Big* could offer an attractive motivation for initiating and solidifying collaboration with others in one's area(s) of scholarly interest.

Less intuitive methods for confirming big. In some cases, a researcher may wish to employ a somewhat less intuitive method for confirming big. Perhaps, for example, they lack the funds and/or collaborative network to use the intuitive methods described above. Alternatively, they may simply desire a more powerful, creative, and/or ecologically valid way of *Confirming Big*. I now briefly discuss some less intuitive methods for *Confirming Big* that possess these characteristics, though, admittedly, they are not feasible for all lines of research.

(Inter)national datasets. Perhaps the most straightforward of the less intuitive methods for psychologists to carry out the *Confirming Big* stage of research, would be for psychologists to test their exploratory models using existing (inter)national datasets. Researchers interested in testing models pertaining to health and wellbeing, for example, might consider *Confirming Big* through the use of the Midlife Development in the United States (MIDUS; http://midus.wisc.edu) or National Longitudinal Study of Adolescent to Adult Health (Add Health; http://cpc.unc.edu) datasets; for models pertaining to political attitudes, researchers could use datasets like the American Values Survey (http://publicreligion.org) or the General Social Survey (GSS; http://norc.org); cross-cultural models, finally, could be investigated through the International Social Survey Programme (ISSP; http://issp.org). In some cases, these datasets may already include all the variables psychologists need to carry out *Confirming Big*, but occasionally they may need to supplement these datasets with coding additional variables related to their hypotheses. If, for example, a researcher was interested in performing a confirmatory test of a hypothesized association between Gelfand and colleagues' (2011) relatively new *tightness-looseness* construct, and some Add Health variables, they might coopt the state-level estimates of *tightness-looseness* provided by Harrington and Gelfand (2014) in order to carry out this analysis.

Google correlate. A relative newcomer to the method toolboxes of social psychologists is Google's Correlate service (https://www.google.com/trends/correlate), in which a researcher can enter weekly or monthly time series data (for a number of countries), or US State-by-Sate data, and Google will return the most frequently searched terms entered by users during that time period (or within each state; see Neville 2012; and Walasek & Brown, 2015 for examples). As such, researchers with temporal and/or geographic corollaries of the processes they are trying to experimentally study may find Google Correlate a particularly informative method of *Confirming Big*. This may be especially true for those attempting to study cultural and/or historical processes. Though, unlike many of the other *Confirming Big* methods, users of Google Correlate will be limited to testing relatively simple models, without a means of including a more diverse array of control variables.

*Social media analysis***.** Another possible avenue for *Confirming Big* is through the collection and analysis of social media data, such as "tweets" from Twitter, or "status updates" from Facebook. Sentiment analysis of social media data is becoming increasingly common (e.g., Schwartz et al., 2013), and could range from utilizing relatively simple pre-programmed psychological language dictionaries (e.g., Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), to complex machine learning/data-driven language categories (Eichstaedt et al., 2015). Whatever the method, sentiment analysis can reveal psychologically rich language related to domains of study like close relationships (e.g., family or friend-related words), construal (e.g., 1st person singular or 1st person plural pronouns), emotion (e.g., positive or negative emotion words), cognition (e.g., certainty or inhibition-related words), and motivation (e.g., death or sex-related words), to name a few. And whereas researchers may be more limited in terms of what variables they can attempt to control for when using Google Correlate, they may be surprised by the amount of additional data that can be coaxed from social media analysis (e.g., sociodemographic data in profiles). Though likely a novel method for many psychologists, there are a number of good resources available that demonstrate how data from social media sites like Twitter and Facebook can be easily collected, such as books by Danneman and Heimann (2014), and Russell (2013), and the *twitteR* package for *R*, by Gentry (2015).

*Meta-reanalysis***.** Finally**,** a relatively novel means for psychologists to conduct strong confirmatory tests of their exploratory models is through a process of meta-*re*analysis. Meta-reanalysis requires scholars to reanalyze an existing large *N* meta-analytic dataset pertaining to their topic, by coding for new, unexplored study-level characteristics, related to their predictions. Scholars would then test whether their newly coded study characteristics moderated effect sizes in the predicted direction. As many social psychological domains already have already been synthesized meta-analytically, meta-reanalysis should prove to be a viable form of *Confirming Big* for a number of research topics (e.g., close relationships, motivation, prejudice, etc.,; see Fraley, 2002; Burke, Martens, & Faucher, 2010; and Sibley & Duckitt, 2008, for respective examples).

What Are the Advantages of *Exploring Small, Confirming Big*?

Exploring Small, Confirming Big would benefit the field in a number of ways, if adopted. Namely, it would help researchers to strike a better balance between exploratory and confirmatory research, increase the level of methodological rigor, meaning, and depth in psychological research, and potentially highlight that *p*-hacking is not necessary when large samples are used in research.

Balancing Exploratory and Confirmatory Research

Adopting an *Exploring Small, Confirming Big* approach to research will provide researchers with a number of benefits throughout the research process. Chief among such benefits will be achieving a more harmonic balance between the exploratory research practices that are thought to be at least partially responsible for replicability concerns (cf. Simmons et al., 2011), and the more strictly confirmatory research approach advanced by Cumming (2014) and others (e.g., Wagenmakers et al., 2012). *Exploring Small, Confirming Big* will therefore leave room for serendipity and exploration in the research process, through the flexibility allowed by the *Exploring Small* stage in which psychologists can explore their data, and reveal unexpected effects (King, 2014).

The *Confirming Big* stage, conversely, should restrict overzealous use of researcher degrees of freedom. Primarily, this will be accomplished through the requirement of preregistration of hypotheses, methods, and analysis plans for *Confirming Big* methods. However, certain *Confirming Big* methods will provide additional restrictions on researcher degrees of freedom. Consider, for example, that a single study may include dozens of different variables that are available as covariates for potential use in *p*-hacking. This is not the case, however, with *Confirming Big* via meta-reanalysis. Meta-analytic datasets typically contain fewer variables for analysis than original studies, and though researchers *could* re-code articles for additional covariates to use in meta-reanalysis, this would dramatically reduce the convenience of *p*-hacking. And finally, although *Confirming Big* through the analysis of (inter)national datasets would allow researchers to draw upon a much larger set of variables, which could facilitate a certain level and type of *p*-hacking, other researcher degrees of freedom are restricted with this approach, such as the fixed sample size (cf. Simmons et al., 2011).

Exploring Small, Confirming Big will ensure researchers feature both exploratory and confirmatory research processes in their scholarship; the two should not be considered antagonistic, but rather, complimentary processes. Further, by permitting the use of NHST, *Exploring Small*, *Confirming Big* could get more buy in from psychologists, by striking a more balanced compromise between NHST and meta-analytic/estimation approaches to research.

Increasing Methodological Rigor, Meaning, and Depth

There are also a number of methodological benefits to be gained by employing an *Exploring Small, Confirming Big* approach to research, particularly within the *Confirming Big* phase. All methods of *Confirming Big*, for example, would result in dramatically increased levels of power for statistically evaluating hypotheses: intuitive approaches (i.e., individual large *N* and Many Lab replications) would achieve a minimum of 80% power; (inter)national datasets often contain thousands of participants, and longitudinal examinations (such as Add Health) only serve to further increase the power of hypothesis testing; social media sites and Google searches offer a seemingly infinite supply of data to analyze; and meta-analysis statistically combines the samples of individual studies included in the analysis. Underpowered studies continue to be widely used throughout psychology (see Maxwell, 2004, for a review). Wide adoption of *Confirming Big* would go a long way towards remedying that trend.

More frequent analyses of social media data would also help to address the concerns expressed by some regarding methodological stagnation in psychological research (cf. Baumeister, Vohs, & Funder, 2007). Though the use of social media data is, by no means, the only way to increase the use of more seemingly meaningful measurement of psychological processes (see Mehl & Conner, 2013; Maner, this issue), it is a method of doing so that is much less resource-intensive (e.g., vs. daily diary studies, or vs. use of invasive technologies).

All four suggested unintuitive forms of *Confirming Big* would also increase the external validity of findings from psychological research. (Inter)national datasets, for example, often rely on random sampling, thereby ensuring generalizable findings based on representative samples.

And though Google search terms, social media data and existing meta-analytic datasets may not be perfectly representative of intended broader populations, they are surely more diverse—and therefore generalizable—than most of the college student or Mechanical Turk-based convenience samples that psychologists often use for their studies.

Finally, the use of meta-*re*analysis when *Confirming Big*, specifically, could reinvigorate a field following an original meta-analysis. As opposed to "closing the book" on a given phenomenon, the publishing of an original meta-analysis would instead pave the way for an number of additional meta-analytic tests of interesting hypotheses. Indeed, social scientists appear to get much better "mileage", so to speak, out of their large (inter)national datasets, which are analyzed by many different research teams, whereas it is more likely to be assumed that original meta-analysts exhaust all theoretically interesting possibilities in their datasets.

Demonstrating That *p*-hacking is Unnecessary

A final benefit of adopting an *Exploring Small*, *Confirming Big* approach is the potential to demonstrate that the need to *p*-hack may be substantively overestimated by scholars utilizing such practices. With substantively increased power in the *Confirming Big* stage, scholars may come to realize that their *p*-hacking efforts (e.g., including a superfluous covariate) are unnecessary to uncover evidence of their hypothesized effects, when large samples are used.

Anticipated Challenges of *Exploring Small, Confirming Big*

Although I anticipate that adopting an *Exploring Small, Confirming Big* approach would benefit our field, doing so would not be without its challenges. Specifically, I expect there to be initial logistical challenges in adopting the less-intuitive methods for *Confirming Big*, as researcher will need to learn the ins-and-outs of analyzing their (inter)national dataset of choice, collecting/analyzing social media data, or conducting meta-analysis. Finally, the use of meta-reanalysis, in particular, will also require several changes to the way meta-analyses are currently conducted by psychologists.

Logistical Challenges

For many, analyzing an available (inter)national dataset, collecting and analyzing social media data, and/or conducting meta-analysis would constitute novel methodology for conducting psychological research. As such, it is likely that there will be initial learning curves to work through, as psychologists acclimate to these less-intuitive way of *Confirming Big*. Analyzing an available (inter)national dataset, for example, may require a psychologist to navigate through the process of requesting access to the data, and becoming familiar with the coding guide for the variables in the dataset. For collecting and analyzing social media data, alternatively, psychologists will need to develop a certain level of coding proficiency in order to access their chosen stream of data, or otherwise become collaborators with computer scientists who may be better equipped to access the desired data. Finally, though psychologists may have a general appreciation for how meta-analyses are conducted and interpreted, it is likely that most do not have formal training in this form of analyses, and may therefore need to invest some time learning about topics such as weighting of effect sizes, coding of study characteristics, and dealing with dependency of effect sizes.

Changes to Meta-Analytic Methods

To maximize the effectiveness of the meta-reanalysis option for the *Confirming Big* stage, a number of changes would need to be made to the way psychologists perform original meta-analyses.

Broadening inclusion criteria for meta-analyses. One of the first changes that would need to be adopted to facilitate meta-reanalysis, would be for original meta-analyses to adopt the broadest inclusion criteria for studies as possible. Many features of studies, which previously might have been used to exclude studies from a meta-analysis (e.g., RCT v. not; published v. unpublished, etc.,), could be potential moderators for future meta-reanalysis. As such, the broadest inclusion criteria should therefore be adopted, and "grey literature" should be thoroughly searched for relevant studies (Rothstein & Hopewell, 2009). Simply stated, if data are attainable, fit the topic at hand, and are interpretable (e.g., reported in a language meta-analysts understand), they should be included in the original meta-analysis.

Ensuring quality of meta-analytic data and reporting. In addition to broadening inclusion criteria, original meta-analyses should go through a much more rigorous review than they do now, particularly with respect to the accuracy of the analyses. Like any other type of quantitative study, calculation and reporting errors can and *do* happen in meta-analysis (e.g., Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007). Since others will be using the same meta-analytic dataset to test their hypotheses via meta-reanalysis, it will be crucial to ensure original calculations and reports are accurate; the impact of calculation and reporting errors will be exponentially compounded by subsequent meta-reanalysis. Original meta-analyses should therefore be subjected to a thorough analytic review (Sakaluk, Williams, & Biernat, 2014) to ensure accurate reporting, as well as verifying correct calculations for a small sample (e.g., 5-10 articles) of effect sizes.

Sharing original meta-analytic data. For meta-reanalysis to be viable, complete meta-analytic data and copies of articles included therein should be available to download (e.g., via the Open Science Framework). As data are values and characteristics of studies—and not people—many of the normal barriers to data sharing (e.g., informed consent, confidentiality, and anonymity of participants) do not apply.

Furthermore, the original meta-analytic dataset should be periodically updated (e.g., once every 1-2 years) with new articles that are applicable; researchers can use search terms, described in the original meta-analysis in their title or keywords, to ensure their articles are included in subsequent updates to initial meta-analyses, or just send their articles to the original meta-analyst directly. Those engaging in meta-reanalysis should also share their new coding with the original meta-analysts, so that these variables may be added to the dataset for others to use in meta-reanalysis as well. All of these goals—sharing and updating of original meta-analyses, and adding moderators from meta-reanalyses—could be greatly facilitated by adopting Tsuji, Bergmann, & Cristia's (2014) proposed community-augmented meta-analysis system.

Incentivizing original meta-analysts to help with meta-reanalysis. Finally, conducting an original meta-analyses requires a *massive* amount of work on behalf of the original researcher(s); my proposed changes to meta-analysis, to facilitate *Confirming Big* via meta-reanalysis, will demand even more (e.g., analytic review, sharing data, etc.,). As such I think it is appropriate to provide incentives for original meta-analysts who are willing to aid in meta-reanalysis.

Original meta-analysts, for example, should be included as the final coauthor(s) on papers including a meta-*re*analysis of their original dataset. If the original meta-analyst(s) does more than provide/describe/orient new researchers to data files and review drafts of manuscripts, they should be considered for higher-level authorship. The original meta-analyst(s) should also be provided a subsection in the General Discussion, in which they may briefly describe any concerns they have with the methods used or conclusions drawn by the new meta-analysts, or otherwise endorse the meta-*re*analysis process and its conclusions.

Some may initially perceive these suggested incentives—especially the bequeathing of co-authorship—as unduly benefiting the original meta-analyst(s). I suspect, however, this concern to be short-lived once, once researchers begin to benefit from the increased confidence that editors and reviewers will have in their work because of the statistical power afforded by a meta-analytic dataset.

Conclusion

Metascientific research on the replicability of psychological science has made clear that change is needed in how psychologists conduct research and analyze their data. *The New Statistics* (Cumming, 2014) is one proposed system that attempts to address these concerns, but many psychologists disagree with some of its larger proscriptions (i.e., banning NHST)—and for good reason. As I have argued, there *are* cases when NHST helps to create a cumulative, replicable science, and it remains to be seen—though I am quite skeptical in most cases—whether small-scale meta-analyses help to advance these disciplinary goals at all.

I have offered the *Exploring Small*, *Confirming Big* approach as an alternative to *The New Statistics* (Cumming, 2014). To be clear, it is intended as an approach for scholars to *consider* for their research, not as a *mandate* for all to follow; there are likely some research questions that are best pursued through some other approach. Even so, *Exploring Small, Confirming Big* addresses some of the key limitations of *The New Statistics*, particularly regarding the banning NHST, and encouraging small-scale meta-analyses; its adoption would provide a number of benefits, and some unique challenge to overcome. In the end, the allowance of NHST, while simultaneously increasing rates of strong confirmatory tests of exploratory hypotheses, may ultimately lead to greater adoption of *Exploring Small, Confirming Big*, compared to *The New Statistics*, and thus, a more cumulative and replicable science of psychology

Footnotes

1. The presence of publication bias can be empirically assessed with a number of NHST-based methods (see Card, 2012, for a review). Based on the outcome of these tests, researchers must then select between competing meta-analytic models (e.g., trim-and-fill v. no trim-and-fill; or the precision effect test v. the precision effect estimate with standard error; see Duval & Tweedie, 2000a, 2000b, and Stanley & Doucouliagos, 2014, respectively).

2. Some available alternatives include: (1) examining the change in the comparative fix index (CFI) between parent and nested models (Cheung and Rensvold, 2002); (2) comparing information criteria, like the Akaike information criterion (AIC) or Bayesian information criterion (BIC), between parent and nested models (Brown, 2015); and (3) using a Monte Carlo simulation methods to compare parent and nested models (Pornprasertmanit, Wu, & Little 2013). The first alternative, however, has only been empirically validated in the context of evaluating factorial invariance; the second alternative lacks established guidelines for what constitutes meaningful amount model fit improvement/degradation; and the third alternative is still being developed and is inaccessible for the bulk of psychologists to use.

3. Though, in spite evidence that researchers misunderstand what NHST results mean or do not mean, psychologists exhibit a similarly high degree of misunderstanding when interpreting confidence intervals (Belia, Fidler, Williams, & Cumming, 2005).

4. I think it is likely—even in the presence of publication bias—that *p*-hacking will bias large-scale meta-analyses to a lesser extent than small-scale meta-analyses, though both will surely be affected by *p*-hacking. It is ultimately, however, an empirical question, and the answer probably depends on the size of the effect(s) in question, the level of assumed publication bias, and the level and frequency of assumed *p*-hacking.

5. I sincerely thank my two reviewers for the very helpful suggestion to include these approaches to *Confirming Big*, in order to increase the broad appeal of the *Exploring Small, Confirming Big* approach.

References

Bakker, M., & Wicherts, J. M. (2011). The (misreporting) of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396-403.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389-396.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd edition). New York, NY: Guilford Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review*, *14*, 155-195.

Carey, B. (November, 2011). *Fraud case seen as a red flag for psychology research*. Retrieved from http://www.nytimes.com.

Chan, M. E., & Avery, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science*, *7*, 79-92.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science*, *334*, 1182.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Danneman, N., & Heimann, R. (2014). *Social media mining with R*. Birmingham, UK: Packt Publishing.

Eich, E. (2014). Business not as usual. *Psychological Science*, *27*, 3-6.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … & Seligman, M. E. P. (2015). Psychological language on twitter predicts country-level heart disease mortality. *Psychological Science*. doi: 10.1177/0956797614557867

Fabrigar, L. R., & Wegener, D. T. (this issue). Conceptualizing and evaluating the replication of research results. *Journal of Experimental and Social Psychology*.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine, and ecology* (Doctoral dissertation). Retrieved from tiny.cc/fionasphd

Fraley, R. C. (2002). Attachment stability from infancy to adulthood: Meta-analysis and dynamic modeling of developmental mechanisms. *Personality and Social Psychology Review*, *6*, 123-151.

Gentry, J. (2015). *Package 'twitteR'*. Retrieved from http://cran.r-project.org/web/packages/twitteR

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *Journal of the American Medical Association*, *298*, 430-437.

Hoyle, R. H (2012). *Handbook of structural equation modeling*. New York, NY: Guilford Press.

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

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.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83-105). Malden, MA: Blackwell.

Klein, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA Books.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, S., Berstein, M. J., … & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142-152.

Kruschke, J. (2014). *Doing Bayesian data analysis, second edition: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

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

Maner, J. K. (this issue). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163.

Mehl, M. R., & Conner, T. S. (2013). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry*, *23*, 291-293.

Neville, L. (2012). Do economic equality and generalized trust inhibit academic dishonesty? Evidence from state-level search-engine queries. *Psychological Science*, *23*, 339-345.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Russell, M. A. (2013). *Mining the social web* (2nd edition). Sebastopol, CA: O'Reilly.

Sakaluk, J. K., Williams, A. J., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, *9*, 652-660.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Argawal, M., … & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*, e73791.

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, *12*, 248-279.

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). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875-1888.

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

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572-582.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stanley, T. D., and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60-78.

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

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Commuity-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*, 661-665.

Uhlmann, E. L., et al. (this issue). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*.

Wagenmakers, E., Wetzels, R., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632-638.

Walasek, L., & Brown, G. D. A. (2015). Income inequality and status seeking: Searching for positional goods in unequal U.S. states. *Psychological Science*, *26*, 527-533.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and quality of reporting of statistical results. *PLoS ONE*, *6*, e26828.