IN PRACTICE

Want to make behavioural research more replicable? Promote single paper meta-analysis

Blakeley B. McShane and **Ulf Böckenholt** argue for single paper meta-analysis to be the default statistical tool whenever multiple similar studies of a common phenomenon are published in one paper

The biomedical and social sciences are facing a widespread crisis of replicability. This problem gained particular recognition in behavioural research fields like psychology and consumer behaviour after numerous prominent findings notoriously failed to replicate. As it turns out, brief exposure to the US flag does not shift support towards the Republican Party for up to eight months, and "power posing" for two minutes does not increase testosterone and decrease cortisol.

While the traditional $p < 0.05$ threshold for statistical significance was once deemed a bulwark against noise-chasing and thus a guarantor of replicability, this conventional wisdom is now under assault. For example, the American Statistical Association released a statement cautioning scientists and policy-makers against basing decisions only on whether a p-value passes a specific threshold.[1] And the literature currently abounds with proposals to redefine or justify statistical significance thresholds – or even abandon them altogether.

Yet a relatively simple technique has, in our view, been underutilised in most publications. That technique is called meta-analysis. While it is commonly used to summarise vast and established literatures, we propose that there is no reason why a single paper meta-analysis (SPM) should not be the default statistical tool whenever multiple similar studies of a common phenomenon are published in a single paper.

Two unique features of behavioural research make this a compelling possibility. First, in an effort to both "self-replicate" and demonstrate robustness to slight differences in the study design, many individual research papers often *do* feature multiple studies of a given phenomenon with such minor variations. Using statistical techniques that can analyse these studies in aggregate seems like a clear win, a no-brainer in terms of improved precision.

Blakeley B. McShane is a statistical methodologist and faculty member in the marketing department of the Kellogg School of Management at Northwestern University.

Ulf Böckenholt is the John D. Gray Professor of Marketing at the Kellogg School of Management at Northwestern University.

Second, in behavioural research it is customarily expected that *each* published study "works", that is, reaches the $p < 0.05$ threshold for statistical significance. But, because individual studies tend to be noisy, screening studies by statistical significance results in published estimates that are biased upwards (often to a surprising degree) and frequently in the wrong direction. For example, one study found that beautiful parents are eight percentage points more likely to have daughters – implausibly larger than the three percentage point effect of extreme conditions, such as famine, found in other research.[2] This inevitably leads to many expensive, demoralising failed replication efforts – a crisis for the field, yes, but also for individual researchers who may spend valuable years chasing fruitless projects. A move towards SPM would allow for a more holistic evaluation of the evidence. Critically, it would offer a path that encourages researchers to publish *all* of their data on a phenomenon – warts and all.

Choice overload

How might this work in practice? Let us consider a researcher interested in writing a paper on the choice overload hypothesis, a popular topic in consumer behaviour. While common sense dictates that more choices should offer consumers a better shot at choosing a product that will ultimately satisfy their needs, a large number of studies over the last 15 years suggest that this is not necessarily the case: when a choice set contains *too many* options, individuals may struggle to make a choice at all, or feel discontent regardless of what they choose.

pbig/Bigstock.com

Given the counter-intuitive and wide-reaching implications of this finding, the choice overload hypothesis has attracted a considerable amount of attention. Researchers have studied choice overload in a host of product categories, using several dependent measures, and examining the effect of several factors that could potentially exacerbate, attenuate, or even reverse it.[3]

TABLE 1 Single-study *p*-values. The column labelled "Contrast 1" gives *p*-values from equal variance *t*-tests of the simple effect of choice set size when there was no time pressure; the column labelled "Contrast 2" gives *p*-values from equal variance *t*-tests of the simple effect of choice set size when there was time pressure; and the column labelled "Contrast 3" gives *p*-values from *t*-tests of the interaction based on the linear model. Only studies 2, 4, 5, 7, and 9 "worked" in terms of yielding statistical (in)significance of all hypothesised comparisons: that is, these are the only studies in which *p*-values failed to achieve statistical significance under contrast 1, while achieving statistical significance under contrasts 2 and 3.

Study	Contrast 1	Contrast 2	Contrast 3
Study 1	0.0183	NA	NA
Study 2	0.5736	NA	NA
Study 3	NA	0.1632	NA
Study 4	NA	0.0018	NA
Study 5	0.7909	0.0000	0.0001
Study 6	0.8544	0.1121	0.3195
Study 7	0.2289	0.0000	0.0000
Study 8	0.2006	0.0004	0.0962
Study 9	0.3868	0.0015	0.0042

Suppose our researcher was interested in studying choice overload in the context of one of these factors – specifically, whether participants presented with large or small choice sets were less satisfied with their choice when they faced time pressure in making the choice as opposed to when they did not. The researcher hypothesised that participants would be less satisfied when choosing from the larger choice set when they faced time pressure, but not when they were given unlimited time to make their decision.

So, the researcher conducted nine randomised, between-subjects studies, each of which varied some aspect of the choice task unrelated to the hypothesis, such as the product category from which participants were asked to choose and whether the decision was made online or in a store. Critically, two studies examined the effect of choice set size when there was no time pressure; two examined the same effect under time pressure; and five looked at all four conditions in order to study the interaction between choice set size and the presence or absence of time pressure. The primary dependent measure in these studies was satisfaction measured on a nine-point integer scale. Summary statistics for these (hypothetical) studies are presented in a supplementary online table – see significancemagazine.com/spm.

Under the status quo, the researcher would analyse each of these studies in isolation. In particular, the researcher would perform (i) separate significance tests of the simple effect of choice set size when there was no time pressure in studies 1–2 and 5–9; (ii) separate significance tests of the simple effect of choice set size when there was time pressure in studies 3–9; and (iii) separate significance tests of the interaction in studies 5–9. In support of the hypothesis, the researcher would hope for all of the first set of tests to fail to reach the $p < 0.05$ threshold for statistical significance, and all of the second and third sets to reach it.

Unfortunately, as shown in Table 1, only studies 2, 4, 5, 7, and 9 "worked" in terms of yielding statistical (in)significance of all hypotheses tested in the study. Thus, the exigencies of the publication process – the complex interaction of editors, reviewers, authors, and other actors – would typically dictate that only these five studies be published.

Instead, it would be useful to look at the results of all nine studies jointly via an SPM, as we report in Figure 1 (page 40) and Table 2. The figure shows that the first simple effect is estimated to be small in each study and nearly zero by the SPM – evidence that participants appear not to be affected by choice overload when they have adequate time to make their choices. The second simple effect, on the other hand, is estimated to be reasonably large in many studies and by the SPM – evidence that when participants are forced to make choices quickly, they *do* experience less satisfaction with choices made from a larger choice set. Finally, while the studies show decidedly mixed statistical significance of the interaction effect, the SPM shows a substantial effect.

In sum, the more comprehensive view of the evidence provided by the SPM yields stronger support for the researcher's hypothesis.

SPM has the additional benefit of quantifying and accounting for heterogeneity (or between-study variation) in effect sizes – something that is possible only by looking across studies and analysing them jointly. This is critical because estimates of standard

TABLE 2 Results of choice overload SPM for all studies and for selected studies (2, 4, 5, 7, and 9 from Table 1). See also Figure 1 (page 40) and singlepapermetaanalysis.com.

Contrast	All studies		Five selected studies	
	Estimate	Std. error	Estimate	Std. error
Large versus small, time pressure absent	–0.0153	0.0708	0.1206	0.0727
Large versus small, time pressure present	–0.4967	0.0708	–0.6311	0.0713
Interaction	–0.4814	0.1001	–0.7517	0.1018
Heterogeneity (I^2 estimate; 95% interval)	52%; 24–70%		0%; 0–17%	

errors that ignore heterogeneity – as single-study estimates do – are optimistically small, yielding (among other things) miscalibrated Type I and Type II error.[4]

In Table 2 (page 39), heterogeneity is quantified via the I^2 statistic, which gives the percentage of the variation in the observations (beyond that attributable to the experimental manipulations) that is due to heterogeneity as opposed to sampling variation. According to norms in behavioural research, the I^2 estimate of 52% (95% CI: 24–70%) denotes a moderate degree of heterogeneity – perhaps not surprising given that the choice task varied from one study to the next. Studies designed to be closer replications of one another would likely yield a lower estimate of heterogeneity but would speak less to the robustness of the phenomenon across such factors as the product category and whether the decision was made online or in a store.

Finally, because SPM shifts the focus away from noisy single-study significance tests and towards a more holistic view of the evidence, it encourages and facilitates full and transparent reporting of *all* studies. This has enormous implications for replicability. For example, Table 2 reports the effect size estimates of the meta-analysis of all nine studies as well as of the selected sample of five studies; as can be seen, the latter leads to inflated estimates of effect sizes and deflated estimates of heterogeneity. Both would in turn lead researchers seeking to follow up on this work to set their sample sizes too low and thus "fail" to replicate these results.

Part of the solution

SPM aids replicability by shifting the focus away from noisy single-study significance tests towards the convergence and variation in estimates across studies and by encouraging full and transparent reporting of all data and results – including study summary statistics (e.g., our supplementary online table). It is also beneficial for theory because the more precise SPM estimates and tests of effects may detect findings that single studies do not and because SPM estimates of heterogeneity can suggest unaccounted-for moderators of theories.

Unfortunately, because behavioural research studies typically feature multiple effects of interest (e.g., two simple effects and the interaction, as in our choice overload example) traditional meta-analytic techniques such as Fisher's method, Stouffer's method, and the dominant standardised effect approach are either unsuitable for SPM or difficult to correctly apply.

To remedy this and facilitate SPM, we have developed methodology that accommodates multiple effects of interest. Our approach is user-friendly because it requires only basic summary information (e.g., means, standard deviations, and sample sizes) and is implemented on an easy-to-use website (singlepapermetaanalysis.com). Despite requiring only basic summary information, the model underlying the methodology is equivalent, by statistical sufficiency, to that underlying the "gold standard" meta-analytic approach – a hierarchical (or multilevel) model fit to the individual-level observations. We direct interested readers to our 2017 paper in the *Journal of Consumer Research* for details.[5]

Given the benefits of this methodology and the ease with which it can be used, we advocate that authors of typical behavioural research papers include a table of summary information from all of their (and potentially others') studies, conduct and discuss an SPM based on this, and display the intuitive graphical summary (e.g., Figure 1). This will supplement the single-study analyses and discussions featured in typical behavioural research papers and requires only a minor modification of current practice.

SPM alone will not solve the replicability crisis in behavioural research. Nonetheless, along with other measures,[6] we believe it can help push researchers away from the pursuit of irrelevant single-study statistical thresholds and the resulting declarations about there being "an effect" or "no effect". This in turn should free their attention to focus on more important concerns for replicability, such as theory, mechanism, and measurement, as well as the estimation of effect sizes, the uncertainty in these estimates, and the variation across them. ■

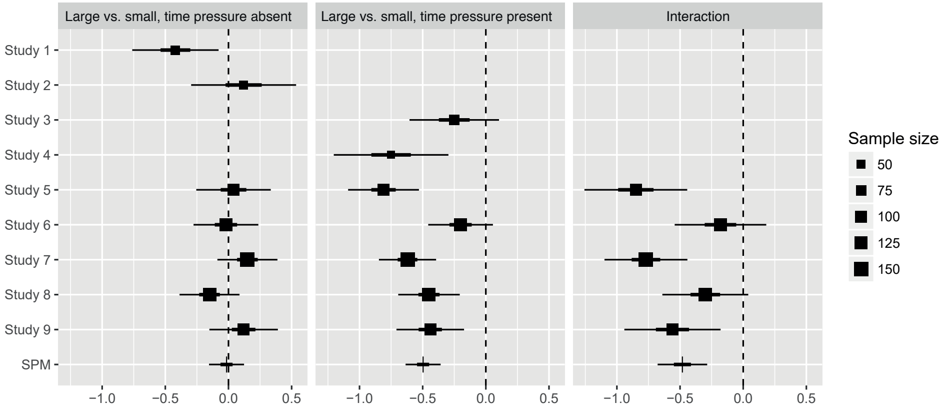

FIGURE 1 Results of choice overload SPM. Effect estimates are given by the squares for single-study estimates and the vertical bars for SPM estimates; 50% and 95% intervals are given by the thick and thin horizontal lines, respectively. Single-study estimates are absent when a study omits a condition relevant for computing a given effect. The average sample size per condition in each study is given by the size of the squares. Effect estimates and estimated standard errors as well as estimates of heterogeneity are in Table 2 (page 39).

Acknowledgement

The authors thank Jessica Love for her valuable advice and suggestions on this article.

References

1. Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician*, **70**(2), 129–133.

2. Gelman, A. and Weakliem, D. (2009) Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, **97**(4), 310–316.

3. McShane, B. B. and Böckenholt, U. (2018) Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, **83**(1), 255–271.

4. McShane, B. B. and Böckenholt, U. (2014) You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, **9**(6), 612–625.

5. McShane, B. B. and Böckenholt, U. (2017) Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research*, **43**(6), 1048–1063.

6. McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2018) Abandon statistical significance. Preprint, arXiv:1709.07588. Forthcoming in *American Statistician*.