Metastudies for robust tests of theory

Beth Baribault[a,1], Chris Donkin[b,1], Daniel R. Little[c], Jennifer S. Trueblood[d], Zita Oravecz[e], Don van Ravenzwaaij[f], Corey N. White[g], Paul De Boeck[h], and Joachim Vandekerckhove[a,1,2]

[a]Cognition and Individual Differences Laboratory, Department of Cognitive Sciences, University of California, Irvine, CA 92697; [b]School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia; [c]Knowledge, Information and Learning Laboratory, Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, VIC 3010, Australia; [d]Department of Psychology, Vanderbilt University, Nashville, TN 37235; [e]Quantitative Developmental Systems Methodology Core, Department of Human Development and Family Studies, Pennsylvania State University, State College, PA 16801; [f]Psychometrics and Statistics, Department of Psychology, University of Groningen, 9712 CP Groningen, The Netherlands; [g]Department of Psychology, Missouri Western State University, St. Joseph, MO 64507; and [h]Quantitative Psychology, Department of Psychology, Ohio State University, Columbus, OH 43210

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved October 3, 2017 (received for review June 29, 2017)

We describe and demonstrate an empirical strategy useful for discovering and replicating empirical effects in psychological science. The method involves the design of a metastudy, in which many independent experimental variables—that may be moderators of an empirical effect—are indiscriminately randomized. Radical randomization yields rich datasets that can be used to test the robustness of an empirical claim to some of the vagaries and idiosyncrasies of experimental protocols and enhances the generalizability of these claims. The strategy is made feasible by advances in hierarchical Bayesian modeling that allow for the pooling of information across unlike experiments and designs and is proposed here as a gold standard for replication research and exploratory research. The practical feasibility of the strategy is demonstrated with a replication of a study on subliminal priming.

robustness | generalizability | metastudy | radical randomization | many labs

Imagine, if you will, an experiment in the psychological laboratory. In the experiment, a single participant provides data in each of two conditions. Further suppose an effect is observed in the form of a mean difference between the two conditions. Unless there are strong reasons to believe that all humans are largely interchangeable with respect to this particular effect, readers and reviewers will reasonably point out that this effect might be idiosyncratic to the participant and hence not generalizable to the broader population.

One potential remedy is for the researcher to replicate the experiment with the same participant and one newly recruited participant—thereby enacting a systematic manipulation of the suspected moderating variable (i.e., participant identity). Such a design enables at least two related claims: possibly that there are individual differences in the magnitude of the effect, and possibly that the effect occurs in some participants but is absent in others.

This strategy is, however, clearly limited: It does not allow for population-level inference. Rather than merely observing that some individual differences could occur, we might instead be interested in whether the effect holds for most humans, or on average across humans, or perhaps for all humans. Such claims call for a hierarchical strategy in which not one or two but many participants are randomly sampled from the population toward which we wish to generalize. If the resultant sample is representative of the population, then the sample mean effect will be an unbiased estimate of the population mean effect and the sample variance in the effect will permit statements about the generality of its occurrence.

In the same way that psychological scientists typically want to generalize from one participant to all potential participants (within certain boundaries), so too will they often want to generalize from a small set of conditions to all conditions (within certain boundaries). For example, researchers who want to claim that stress impairs memory presumably believe that this effect is not specific to the particular aspects of one specific experiment. However, testing the myriad experimental "facets," or moderators, involved (e.g., setting, stimuli, etc.) can be burdensome, time-consuming, and expensive. The strategy of random selection is a sound and viable one for potential moderators of an experimental effect, including potential moderators other than participant identity. In particular, we believe that extensive randomization can lead to scientific conclusions that are more general in scope, more robust to incidental variations in experimental setup, and more likely to replicate in future studies.

In what follows, we will we introduce the concept of a "metastudy," in which we combine "radical randomization" (RR) of experimental features and systematic pooling of information with a Bayesian hierarchical model. We argue that sampling from a population of possible experiments in the same way one would sample from a population of possible participants is a practically feasible approach that can increase the robustness of empirical findings in psychology.

Causes of Nonreplication and Variations on Replication. Replicability of empirical findings has been a central topic in recent psychological science. Following a series of dramatic revelations in which researchers have appeared unable to reliably replicate empirical effects once thought to be robust, there is now talk of a "crisis of confidence" (1) in the field. While there are a number of possible explanations for the lack of replicability (2), one commonly indicated problem is the issue of publication bias: the preference to publish statistically significant results (i.e., results that lead to the rejection of a null hypotheses; refs. 3 and 4). This statistical significance filter (5) biases the published record toward results that capitalize on measurement noise and fluke outcomes (6).

Moreover, evidence from psychological studies—even if published without bias toward certain outcomes—is often weak due to traditions of insufficient sample sizes and noisy measurement tools, which lead to generally low ability to detect true effects and a concomitant increase in false-positive results (7, 8). The combination of publication bias and low standards of evidence

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: B.B., C.D., and J.V. designed research; B.B., C.D., D.R.L., J.S.T., D.v.R., C.N.W., and J.V. performed research; Z.O. and J.V. analyzed data; and B.B., C.D., D.R.L., J.S.T., Z.O., D.v.R., C.N.W., P.D.B., and J.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All materials and data are freely available online at https://osf.io/u2vwa/.

[1]B.B., C.D., and J.V. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: joachim+pnas@uci.edu.

Published online March 12, 2018.

would naturally cause frequent failures to replicate, since effects claimed in the published literature are likely to be false alarms. Given the uncertain nature of one-off effects found in the literature, replication of empirical results is a clear gold standard of convincing evidence: Greater confidence is warranted in theories whose predictions repeatedly come true (9) or whose predictions survive repeated falsification attempts (10).

At the same time, even when a published effect is true, it is possible for effects to fail to replicate strictly due to seemingly innocuous differences in the implementation of the experiment (i.e., due to "hidden moderators" that may occur in replication studies). Small variations in experiments are of course unavoidable: Exact replication is strictly impossible. However, for the purposes of creating generalizable knowledge what matters most is recreating the necessary and sufficient conditions that will show the effect as predicted by some theory. By implication, small experiment variations that are not theoretically relevant should have only minimal impact on the size of a true effect. Indeed, theoretical statements made by researchers almost without fail imply some degree of robustness to irrelevant variables. It was recently proposed that authors make these claims explicit as part of every paper (11, 12).

Such robustness is, of course, a testable assertion. We could take any one of these suspected hidden moderators, systematically vary it as an independent variable (IV) in an experiment, and quantify any differences so obtained. Much theoretical knowledge grows exactly in this fashion.

A related distinction that is often made among replication attempts is that between direct and conceptual replications. A direct replication is one in which the replicating team attempts to follow the original protocol as closely as possible, allowing for no moderating variables that might distort the findings or obfuscate the effect seen in the original publication. In a direct replication, the exact same theoretical prediction—that is, the same hypothesis—is tested. A conceptual replication, conversely, is one in which the replicating team tests the same theory but uses a different instantiation of theory to hypothesis, with entirely different values on some independent variables and possibly different dependent and independent variables as well. In such a replication the issue at hand is the robustness of a reported effect to theoretically irrelevant design variations.

Both of these approaches have associated problems. A common concern about direct replications is that it is typically impossible to copy a protocol exactly: Replications tend to take place at a different time and place from the original, with different subjects, and they are often by a different laboratory with slightly different ineffable and undocumented practices, and not all of the relevant details are reported in the original publication. Conceptual replications, however, lack falsification power: A lack of effect may be due to one of the many differences between the original and the replication. While irrelevant within the adopted theoretical framework, an innocuous difference in design might in fact be a genuine moderating factor. As such, the masking of an otherwise replicable effect by a hidden moderator and a genuine failure to replicate are strictly unable to be teased apart with conventional techniques.

Radical Randomization

Here we present an alternative take on replication that involves the RR of many features of an experiment. As an example, imagine a study in which researchers are interested in some difference between two manners of stimulus presentation. A visual stimulus (e.g., the symbol v) is either presented to the participant normally for a short time (e.g., 30 ms) or it is presented with temporal masking—meaning that it is preceded and followed by visual masks (e.g., strings of symbols such as &&&). These masks are called "forward masks" and "backward masks," respectively, and their addition sometimes suppresses the conscious perception of the temporally flanked stimulus. Such an experiment has a few immutable features that are necessary to address the question at hand (critically, some stimuli need to be masked while others are not). However, many of the features of this experiment are chosen largely arbitrarily: Presumably there is nothing special about the symbol v and the same differences could be illustrated with the symbol b instead, and presumably ### is as effective a forward or backward mask as &&&. If the effect exists, it should shine through—if perhaps diminished—for many different symbols and many different small variations on the experimental setup.

In an RR design, this presumption of robustness is put to a critical test. Rather than consistently using the symbol v, we instead randomly choose any symbol from a set and then choose a new symbol whenever we can (without harming the validity of the study). Such a design could be considered defensive in the sense that it hardens our conclusions against minor infidelities in future replication attempts (i.e., replication attempts that are not strictly faithful and hence are not direct replications)—infidelities such as using a different symbol. That is, the RR design makes conclusions more robust because it mimics some of the potential variance between an experiment and future replication attempts that are—as all replications are—inexact.

To distinguish those immutable IVs that are needed to define the effect of interest from the innocuous design features (strictly speaking also IVs) that are randomized, it will be useful to introduce some new terminology. Borrowing from generalizability theory (13), we call these to-be-randomized IVs facets, and we call a study with many facets a metastudy. While a typical IV has a limited set of values that we normally call conditions, the values of a facet are drawn randomly from a potentially infinite population. We call the values of a facet that happened to be drawn for a particular metastudy its "levels," and we call each cell in the multifaceted design a "microexperiment." The immutable IVs that occur in each microexperiments will be called "elementary IVs." Finally, it will sometimes be useful to think of the population of possible microexperiments, which is defined by the space spanned by all of the facets of a study. We call this the "method space."

Facets can be simple design choices (e.g., the exact stimuli selected from a larger pool), natural constraints (e.g., the geographical location of the laboratory), or explicitly labeled nuisance variables that are randomized (e.g., individual differences between participants). The goal of introducing variability in a facet is to investigate the generality of an effect within a much broader subspace of the method space than is commonly the case. If an effect remains, despite variability in some design features, we establish robustness: invariance of the effect to reasonable variation in the facet. Alternatively, the effect may turn out be sensitive to such variability.

What constitutes "reasonable variation"—as formalized by the distribution from which levels of a facet are drawn—is up to the judgment of the researcher. The sampling distribution of a facet determines the "universe of intended generalization": the range within which we aim to establish the existence of the effect. In general, levels should be sampled so that they well represent the range of the facet across which one hopes to draw conclusions.

Facets may be of particular interest when they are predicted—by one theory or another—to moderate an empirical effect. In such cases, establishing the moderating influence or the invariance of the effect are both of theoretical interest. However, the purpose of an RR procedure is not to build or refine theories as much as it is to establish that an effect holds. Researchers setting up a metastudy are therefore recommended to be liberal in which facets they select for randomization.

We are of course not the first to suggest randomization of experimental features. Indeed, in 1973 psycholinguist H. H. Clark (14) suggested it as a treatment for what he called the language-as-fixed-effect fallacy, and R. A. Fisher (9) famously

proposed it to avoid systematic effects of sampling locations in agricultural experiments. Our position might be characterized as an objection to a broader error of inappropriate use of fixed effects.

Finally, we should point out that randomization itself is not unique in its suitability toward the goal of obtaining a representative sample (15). We merely propose it here as a convenient practical approach to exploring the space of possible microexperiments.

Individually Weak, Jointly Powerful. The RR approach that we propose involves the implicit construction of many microexperiments and randomly sampling among them. A microexperiment might consist of all of the trials that share a level of one selected facet (hundreds or thousands of trials) but may be as small as all of the trials in a single block by a participant (a few dozen trials). What constitutes a microexperiment is less a design decision than a feature of the statistical analysis: It is a grouping of observations that is homogeneous in the facet(s) of interest (but has variability in the elementary IVs so that contrasts can be computed).

Individually, these microexperiments do not deliver much evidence for or against the existence of an effect. However, a key component of the approach is the use of modern statistical techniques (e.g., Bayesian hierarchical modeling and meta-analysis; refs. 16 and 17) to pool information across datasets efficiently.

Theory Testing. A metastudy serves to make a stronger statement about the existence of an empirical effect—namely, its persistence across variations on an experiment. To test an effect in such a hierarchical scenario it is more beneficial to increase the number of independent variations than it is to increase the number of data points. Hence, by randomly sampling many locations in the method space and conducting a small independent experiment in each location the multifaceted design allows robust and statistically powerful statements about the effect.

A theory, with a universe of intended generalizability, can be formalized as an effect size function over a region within a method space—rather than over a point, which would represent a more local hypothesis. The region of the method space within which an effect presents itself allows us to make empirically backed statements about the constraints on generality—that is, the boundary conditions of the theory—that are usually only implicit in psychological theories.

While this strategy seems straightforward—perhaps even obvious—it is to the best of the authors' knowledge essentially unused in psychological or cognitive science. Over time, research groups with a concerted study program eventually develop a portfolio of experiments that vary in small ways, and in that sense these groups work to establish robustness (or observe the lack of it). However, the systematic execution of such a population of experiments—in what we here call a metastudy—does not occur, leading to the potential for bias and correlated error. We believe that the multifaceted design has great potential as a defensive design strategy that allows for more general statements and tests of theory and is likely to yield conclusions that are more robust to small variations in design implementation.

Statistical Analysis of Multifaceted Designs

The multifaceted design affords a number of different statistical approaches. In this section we discuss three possibilities. In the case example, we will demonstrate all three.

In what follows, we will assume an experimental metastudy with some set of elementary independent variables that are theoretically interesting (i.e., whose effect on a dependent variable we are hoping to quantify) and some set of facets. Most facets are not relevant according to the theory we are testing but might be relevant according to some unspecified rival theory or be relevant in ways that are simply not yet discovered.

Global Tests. Many experimental studies are specifically designed to answer a particular question, often of the unary form "is A different from 0?" or the binary form "is B greater than C?" Even though we often have multiple, randomly selected participants and we expect there to be person-level variability, the random effect of participant identity is often ignored on the (reasonable) assumption that with a sufficiently large sample any interindividual differences will "wash out" so the sample is balanced and the sample mean effect is a good estimate of the population mean effect. With the same argument, we can—in a first pass—ignore the differences between the randomly sampled levels of the facets in an experiment. This way, we are able to test for the existence of an inequality on average over the range of possible values of the facet.

The formulation of the model is somewhat standard. Letting $y_{m(i)}$ stand for the dependent variable observed at trial i (which is nested in microexperiment m) and letting $x_{km(i)}$ stand for for the corresponding value of the kth elementary IV X_k (where conventionally $x_{0m(i)} = 1$ to represent the intercept), the global test model has a set of regression weights β_k and a variance ς^2. Errors $\epsilon_{m(i)}$ are identical and independently distributed (i.i.d.) standard normal:

$$y_{m(i)} = \sum_{k=0}^{K} \beta_k x_{km(i)} + \varsigma \epsilon_{m(i)}.$$

This fairly common formulation subsumes as special cases the models associated with the t test (if $K = 1$ and X_1 is binary), linear regression (if X are continuous), or ANOVA (if $K > 1$ and all X_k are binary).

We emphasize, however, that such a global test is only valid if the results are relatively homogeneous between microexperiments. In the same way that ignoring large individual differences may invalidate the results of a conventional experiment, if a facet causes true heterogeneity in the effect size the global test can be a poor approximation, and it is important to evaluate whether the test is appropriate before drawing conclusions from it.

Level-2 Heterogeneity and Moderation. Experimental effect sizes are inherently unstable. Even in the absence of explicit moderators any set of experiments will show variance even in the true effect size—that is, above and beyond measurement error. This instability—which occurs due to ephemeral differences even between superficially identical designs—is sometimes referred to as level-2 heterogeneity.

The global hypothesis test above makes no statement about the robustness of the finding to variations in the experimental setup. To evaluate robustness we can apply a hierarchical model in which a facet is allowed to interact with any or all of the elementary IVs (including the intercept). We then inspect if and how the effect varies over the range of each of the individual facets. In the hierarchical model the regression weights are decomposed to yield the following random-effects model equation:

$$y_{m(i)} = \sum_{k=0}^{K} \left(\beta_k + \sigma_k \gamma_{km}\right) x_{km(i)} + \varsigma \epsilon_{m(i)}.$$

Here, the new parameter γ_{km} indicates the unique contribution of the facet to the effect of the kth elementary IV. The parameter is i.i.d. standard normal. Of primary interest in this scenario is σ_k, the level-2 variance of the contribution of the facet to the effect size β_k, and potentially the pattern of change in γ_{km} across its levels m. The former quantifies the heterogeneity of the effect size: σ_k can be compared with the fixed effect size β_k for reference;

the ratio $\rho_k = \sigma_k/\beta_k$ is sometimes called the coefficient of variation. The parameter ρ_k may be interpreted as a measure of robustness, with small values (say, less than 1/3 or 1/4) indicating robustness and large values indicating sensitivity to the facet k. The changes in γ_{km} over the facet allow us to visualize and study its influence.

While it is sometimes sufficient to visualize an effect or a pattern of effects across values of a moderator, we occasionally need to test whether an effect is nominally present or absent in a given condition. For this purpose, we can use a Bayes factor (or likelihood ratio), which expresses by how much the relative probability of a pair of hypotheses changes when the data are taken into account. That is, if $\mathcal{H}_a$ and $\mathcal{H}_b$ are the hypotheses under consideration and x is the data, the Bayes factor is given by

$$B_{ab} = \frac{P(\mathcal{H}_a|x)/P(\mathcal{H}_b|x)}{P(\mathcal{H}_a)/P(\mathcal{H}_b)}.$$

We will interpret $B_{ab} \geq 10$ as strong support for $\mathcal{H}_a$.

Planned Meta-Analysis. A metastudy will typically lead to somewhat larger datasets than are common in psychological science. To apply a high-dimensional statistical model to a large dataset we use one particularly useful approximation that changes our analysis from a standard hierarchical model into a planned meta-analysis. The approximation is based on the central limit theorem, which allows us to substitute n_m normally distributed data points $y_{m(i)}$ with variance ς^2 by their means $\bar{y}_m$ with SD equal to the SE of measurement s_m:

$$\bar{y}_m = \sum_{k=0}^{K} (\beta_k + \sigma_k \gamma_{km})\, x_{km} + s_m \epsilon_m.$$

A conventional meta-analysis involves a set of studies, each of which can be represented as a point in the method space, with the exact location chosen by the experimenters. The meta-analyst then computes a weighted average of effect sizes across these studies. While conventional meta-analysis is often plagued by severe issues such as publication bias, this is not a concern for the metastudy. Similarly, the issue of hidden moderators is reduced here since at least some differences between microexperiments are recorded: Facets are explicitly identified and their levels are not arbitrarily chosen but—to the extent possible—fairly and independently sampled from a well-defined population distribution.

In the following section we will apply these methods and analyses to an experimental study in cognitive science. For the purposes of exposition we will omit some detail regarding the experiment (full details are available at https://osf.io/u2vwa/).

The Effect of Masked Cues on Cognitive Control

As a toy demonstration, we replicate a recently published experiment in cognitive psychology. Reuss et al. (ref. 18; see especially figure 1) describe an experiment in which a cue that is presented for a subliminal amount of time (i.e., too briefly to be consciously detected) influences how participants balance speed and accuracy in a response-time task. This design has obvious facets (e.g., the color of the cue) whose exact values are not expected to affect the finding of subliminal perception: If the effect is robust, it should appear at all values of the facet; if it is fickle, it should appear in some (contiguous) value ranges but not in others; if it is false, it should not consistently appear in any range of values.

The Basic Task. In the experiment participants were shown a "bullseye" stimulus consisting of a dot surrounded by nine concentric circles. The stimulus appeared either in the right or the left half of the screen and participants were instructed to move the mouse pointer from the center of the screen to the center of the bullseye and then click the left mouse button. Shortly before the presentation of the stimulus a single-letter cue was presented, instructing participants to either favor accuracy (measured in distance from the center) or favor speed. Additionally, the cue was either masked (by the rapid presentation of two three-symbol strings like ### and &&&) or not, giving rise to four experimental conditions. Of primary interest is the effect of the masked cue instruction on the speed and accuracy of the responses that Reuss et al. (18) first reported.

Sampling the Method Space. During the development of the study the experimenters collaboratively constructed a list of facets to include. In Table 1 we list facets related to timing, including the duration of the first and second forward mask, of the first and second backward mask, of the masked and unmasked cue; facets related to color, including the hue and luminance of masks and cues; and other miscellaneous facets, such as the symbols used in the mask and the testing location.

Each of these facets was assigned a distribution from which its values were to be randomly sampled at the beginning of each microexperiment. In almost all cases this involved a uniform distribution over a range of integer values (e.g., the variables relating to presentation time were naturally expressed as an integer number of frames). For one facet, variance was introduced not through random sampling but by a convenience sample: The experiment was conducted in six different geographical locations.

The Experiment. Each participant's session of the experiment began with 16 practice trials whose facets were set to match the original study by Reuss et al. (18) as closely as possible. After that, each block of trials consisted of (*i*) 40 "bullseye" trials whose facets were set to a random value sampled from the corresponding distribution and (*ii*) 40 "cue identification" trials whose facets were set to the same values used in the immediately preceding bullseye block. The first eight trials of each type were considered practice trials as well. The goal of the cue identification trials was to confirm the true subliminal nature of the masked cue. Crucially, all facets' values were resampled at the start of

Table 1. Heterogeneity over facets

Facet	Levels	Original	$\hat{\rho}^*$
First forward mask duration	0–50 ms	40 ms	0.42
Second forward mask duration	0–50 ms	30 ms	0.52
Total forward mask duration	0–100 ms	70 ms	0.59
First backward mask duration	0–50 ms	40 ms	0.49
Second backward mask duration	0–50 ms	30 ms	0.52
Total backward mask duration	0–100 ms	70 ms	0.69
Masked cue duration	0–50 ms	30 ms	0.90
Blank interval duration	250–750 ms	500 ms	0.93
Intertrial interval duration	500–1,500 ms	1,000 ms	0.55
Mask and cue color	13 colors†	White	0.13
Mask and cue contrast	$0.5 \leq \times \leq 1.0$	1.0	0.21
Target center color	13 colors†	Red	0.10
Target center contrast	$0.5 \leq \times \leq 1.0$	1.0	0.21
Target surround contrast	$0.5 \leq \times \leq 1.0$	1.0	0.22
First mask symbol	@, #, $, %, &, ?	#	0.06
Second mask symbol	@, #, $, %, &, ?	%	0.06
Location	Six locations‡		0.36

*$\hat{\rho}$ indicates the observed heterogeneity that the facet introduces in the effect of masked cues on accuracy (lower values indicate greater robustness).

†Twelve hues were sampled between integer multiples of 30° angles in HSV color space; the 13th color was white.

‡The locations were the research laboratories of authors C.D. (Sydney, Australia), C.N.W. (Syracuse, NY), D.R.L. (Melbourne, Australia), D.v.R. (Groningen, The Netherlands), J.S.T. (Nashville, TN), and J.V. (Irvine, CA).

each block of bullseye trials, making each block of trials a unique microexperiment.

Practice trials were discarded. At each bullseye trial we recorded two dependent variables: (*i*) the participant's response time and (*ii*) the distance (in millimeters) between the center of the stimulus and the point where they clicked. In the cue identification trials we recorded (*i*) the response time and (*ii*) the (binary) accuracy. We discarded trials where the reaction time was too high (over 2,500 ms) or too low (under 150 ms) and where the participant clicked without moving the pointer.

Each of the six participating laboratories decided how many blocks each participant would complete (all laboratories chose 14 blocks, which made for ∼1-h sessions) and how many participants would be recruited; with no fixed stopping rule set. Laboratories recruited between 47 and 78 participants from their institutional human subjects pools, for a total of 346 participants and up to 4,844 microexperiments, all with randomly drawn levels on each facet.

The experiment was approved by the institutional review boards of University of California, Irvine (2015-1802), Syracuse University (13-269), Vanderbilt University (151563), University of Groningen (15122-NE), University of New South Wales (153-2387), and the Melbourne School of Psychology (1544198.3). All participants provided informed consent at the beginning of the experiment and were informed that participation was voluntary.

The Dependent Variable. Throughout the following analyses, the quantity of interest is the magnitude of the conditional effect of the cue when it is masked—that is, the difference between the masked-cue, accuracy-instruction condition and the masked-cue, speed-instruction condition. For the purposes of exposition, we will focus only on the dependent variable "accuracy" (negatively coded as the distance from the center of the bullseye), but similar results were found for the "reaction time" dependent variable.

Level-2 Variability. To quantify the heterogeneity between the 4,844 microexperiments we applied a hierarchical Bayesian model (19) that included a unique effect size parameter for each microexperiment (i.e., a random effect of microexperiment). This results in a distribution of effect sizes with as many values as there were microexperiments. Focusing on the effect of masked cues only, the mean of that effect size distribution was estimated at $\hat{\beta} \approx 3.36$ mm. However, its population SD was $\hat{\sigma} \approx 6.46$ mm and the coefficient of variability was $\hat{\rho} \approx 2$, which indicates that the effect is sufficiently sensitive to the differences between microexperiments that it will occasionally vanish.

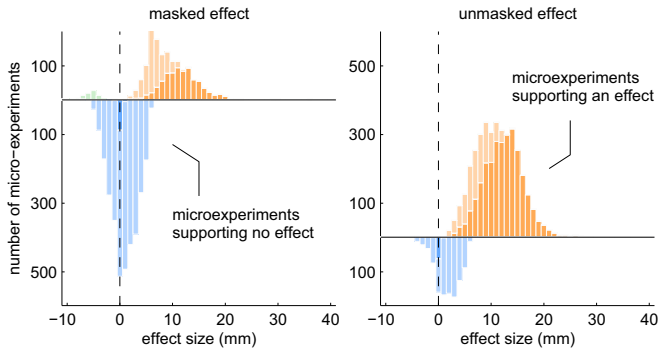

Fig. 1. Level-2 variability. Histograms of estimated effect sizes across microexperiments are split between masked (*Left*) and unmasked (*Right*) conditions and between microexperiments that support an effect (regular bars) versus no effect (inverted bars). Darker bars indicate stronger support with a Bayes factor of at least 10. A majority of microexperiments show support for the unmasked effect, but a similarly large number support no effect of the masked cue.

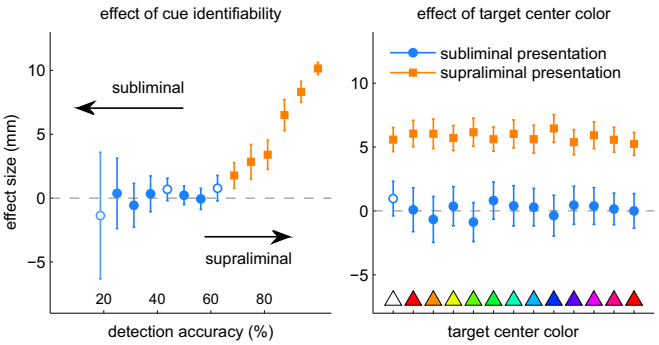

Fig. 2. (*Left*) Microexperiments support an effect when participants are able to consciously identify the cue (square markers), but not otherwise (round markers). (*Right*) The data are split by subliminality. The facet "target center color" was varied over 13 possible levels, but the facet does not appear to moderate the effect of interest. That is, the effect appears robust against this facet. In both panels error bars show 99% credibility intervals. Solid square markers indicate strong evidence (Bayes factor >10) for a nonzero value. Solid round markers indicate strong evidence for a zero value. Empty markers indicate ambiguous evidence.

A histogram of the distribution of effect sizes over microexperiments (Fig. 1) shows the large variability. To construct these histograms, we computed Bayes factors* to express the statistical support for a nonzero effect in each microexperiment. The sample effect sizes more consistent with a zero effect make up the inverted histogram. The figure shows that three-quarters of the individual microexperiments in the masked condition appear more consistent with no effect than with a positive effect and a small number show an effect in the opposite direction. By contrast, in the unmasked condition the large majority of microexperiments are more consistent with a positive effect.

The large variability appears to suggest the existence of one or more moderating variables hidden in our design. We can quantify the heterogeneity of this effect by applying a sequence of hierarchical models. In each model we will estimate the variability of the effect size across levels of one facet (i.e., a random effect of the facet). Each such analysis will yield an estimated coefficient of variability associated with that facet. These estimates are given in Table 1. The largest heterogeneity is seen in the various timing facets, and the effect is particularly unstable across levels of "masked cue duration" and "blank interval duration," while it appears to be relatively robust to changes in colors and symbols.

Moderator Analysis. The observed heterogeneity can be explored by the explicit introduction of potential moderators of the effect. One candidate moderator that is not included in Table 1 is the subliminality of the cue as presented. Recall that after each block of bullseye trials participants completed a block of trials in which they were asked only to identify the cue. In these cue-identification blocks, the cue was presented with the same settings (i.e., the same values on the relevant facets) as in the

*The Bayes factors express how much less likely the effect size of 0 mm is under its posterior distribution than under its prior distribution. The prior distribution of the effect size $\hat{\beta}$ is derived from the prior distributions of the condition means, which was in turn derived from the source paper (18). Assuming a repeated measures correlation of no more than 0.5, the effect size prior worked out to a normal distribution with mean 0 mm and SD 10 mm. This test is maximally sensitive to effect sizes that are slightly smaller than the global mean effect size in the original paper. None of our conclusions regarding Fig. 1 is sensitive to reasonable variation in these assumptions.

bullseye trials. We can quantify the subliminality of the cue under these conditions by the accuracy in the cue-identification trials.

Fig. 2, *Left* shows how the effect of the masked cue varies as a function of the subliminality of the cue presentation. Only in those microexperiments where the cue identification accuracy is at least 68% does an effect of the masked cue appear. In the figure, square markers are filled if the data strongly support an effect (with a Bayes factor of at least 10), round markers are filled if an effect size of zero is strongly supported, and empty markers indicate ambiguity. Each facet can be explored in a similar way to evaluate whether it moderates the effect of interest.

The level-2 variability analysis hinted at the presence of a potential moderator, and Fig. 2 identifies subliminality as one. We can construct similar figures to indicate the lack of a systematic effect of a facet. For example, a facet that is an unlikely moderator is the color of the target center. In Fig. 2, *Right*, we graph the effect size as a function of this facet, splitting microexperiments according to whether the cues were consciously visible. The effect appears to be robust to changes in this facet since it occurs across all levels of the facet for supraliminal trials (squares) and nowhere for subliminal trials (circles).

Conclusion. The effect of masked cues is strongly qualified by the moderator analysis. Masked cues seem to have an effect on participant behavior only in those settings where the cue is consciously visible. We find no evidence of an effect of subliminally presented cues. On the contrary, our data are more consistent with no effect when the cue presentation is truly subliminal.

Discussion

Robustness and generalizability of empirical results are critical considerations regarding the reproducibility crisis that has beset psychological science. The RR approach to experimental design, in which features of an experimental design are strategically randomized, allows researchers to make statements that are less sensitive to unavoidable between-study variability. When a single experiment demonstrates the existence of some effect there is the risk that the effect is isolated to a particular "sweet spot" in the method space. By contrast, the metastudy allows us to make statements about effects in regions in a method space: a well-defined and formalized universe of intended generalization.

In our view, metastudies complement the standard approach to empirical research. The RR approach speaks to the robustness of empirical effects, but such information is only useful to the extent that it informs the development of substantive theory. Experiments with tight control and fixed effects are an established means of generating theoretical explanations for data; we view metastudies as an efficient way of testing such theories by complementing the fixed effect approach with random effects.

The strategy has some weaknesses to keep in mind. First, it is impractical in certain settings, such as when data are expensive to collect. However, it is particularly well suited for "many labs"-style projects in which an ad hoc consortium of research laboratories collaborates in data collection. Still, a metastudy could very reasonably be run within a single laboratory—from a logistical standpoint, the cost to each laboratory that contributed to the applied example was comparable to that of a typical experiment in cognitive science (arguably it was slightly lower since the study materials were produced entirely by the University of California, Irvine and University of New South Wales laboratories). Second, in all but some cases it will be impossible for a research team to identify all facets that might moderate an effect. It serves to remember that claims of generality remain confined to the actually realized method space. However, the randomization of experimental features does provide for a built-in test of some robustness to small variations in experimental features, it can be used to spot weaknesses in an experimental design as well as in empirical claims, and it can be used to generate novel hypotheses when a facet unexpectedly turns out to be influential.

The major strength of RR, and the reason why we recommend it, is that it allows for defensive design: a design strategy under which studies are optimized for generalizability, replicability, and robustness.

ACKNOWLEDGMENTS. This work was supported by National Science Foundation Grants 1230118 and 1534472 (to J.V. and B.B.) and Australian Research Council Grants DP160102360 (to D.R.L.) and DE130100129 (to C.D.).

1. Pashler H, Wagenmakers EJ (2012) Editor's introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspect Psychol Sci* 7: 528–530.
2. Francis G (2012) Publication bias and the failure of replication in experimental psychology. *Psychon Bull Rev* 19:975–991.
3. Guan M, Vandekerckhove J (2016) A Bayesian approach to mitigation of publication bias. *Psychon Bull Rev* 23:74–86.
4. Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychol Bull* 86:638–641.
5. Vasishth S, Gelman A (2017) The illusion of power: How the statistical significance filter leads to overconfident expectations of replicability. arXiv:1702.00556.
6. Sterling TD (1959) Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *J Am Stat Assoc* 54:30–34.
7. Etz A, Vandekerckhove J (2016) A Bayesian perspective on the reproducibility project: Psychology. *PLoS One* 11:e0149794.
8. Gelman A, Loken E (2014) The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am Scientist* 102:460–465.
9. Fisher RA (1935) *The Design of Experiments* (Oliver and Boyd, Edinburgh).
10. Popper K (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge* (Routledge & Kegan Paul, Abingdon, UK).
11. Simons DJ, Shoda Y, Lindsay DS (in press) Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect Psychol Sci*, 10.1177/1745691617708630.
12. Kenett RS, Rubinstein AA (2017) A generalization approach to reproducibility claims. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035070.
13. Cronbach LJ, Rajaratnam N, Gleser GC (1963) Theory of generalizability: A liberalization of reliability theory. *Br J Stat Psychol* 16:137–163.
14. Clark HH (1973) The language–as–fixed–effect fallacy: A critique of language statistics in psychological research. *J Verbal Learn Verbal Behav* 12:335–359.
15. Worrall J (2007) Why there's no cause to randomize. *Br J Philos Sci* 58:451–488.
16. Gelman A, Hill J (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge Univ Press, Cambridge, UK).
17. Sutton AJ, Abrams KR (2001) Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 10:277–303.
18. Reuss H, Kiesel A, Kunde W (2015) Adjustments of response speed and accuracy to unconscious cues. *Cognition* 134:57–62.
19. Stan Development Team (2014) Stan: A c++ library for probability and sampling. Available at mc-stan.org/.