Psychonomic Bulletin & Review
2009, 16 (4), 617-640
doi:10.3758/PBR.16.4.617

THEORETICAL AND REVIEW ARTICLES

What is the probability of replicating a statistically significant effect?

JEFF MILLER
University of Otago, Dunedin, New Zealand

If an initial experiment produces a statistically significant effect, what is the probability that this effect will be replicated in a follow-up experiment? I argue that this seemingly fundamental question can be interpreted in two very different ways and that its answer is, in practice, virtually unknowable under either interpretation. Although the data from an initial experiment can be used to estimate one type of replication probability, this estimate will rarely be precise enough to be of any use. The other type of replication probability is also unknowable, because it depends on unknown aspects of the research context. Thus, although it would be nice to know the probability of replicating a significant effect, researchers must accept the fact that they generally cannot determine this information, whichever type of replication probability they seek.

Scientific theories are built on replicable phenomena (see, e.g., Falk, 1998; Guttman, 1977; Tukey, 1969; Wainer & Robinson, 2003). In sciences with deterministic measurements, the idea of replication is simple: If two researchers measure the same phenomenon using the same instruments and procedures, they should obtain essentially the same results. Things are not so simple when the measurements are subject to random variability due to measurement error, individual differences, or both. In this case, real effects are only replicated with a certain probability—often called the "replication probability." Even when a real effect is present, some replication failures must be expected as one of the unfortunate consequences of variability.

For researchers faced with random variability, it is useful to understand the nature and determinants of replication probability for at least three reasons. First, this probability is relevant in assessing the implications of discrepant results ("Is this a real effect that by chance was not replicated, or was the initial finding spurious?"). Second, it is also relevant when researchers want to show that an effect obtained in one circumstance disappears in some other situation (e.g., a control experiment); the absence of the effect in the new situation is only diagnostic if the experiment had a high probability of replicating a true effect. Third, replication probability is relevant when planning a series of experiments ("What are the chances that I will obtain this effect again in future experiments like this one?").

Unfortunately, there is evidence that many psychological researchers do not understand replication probability (see, e.g., Tversky & Kahneman, 1971). For example, in an oft-cited (e.g., Cohen, 1994) study, Oakes (1986) presented a group of 70 researchers with a scenario in which a two-group comparison resulted in a t test that was significant at the level of $p = .01$. A majority (60%) thought that this indicated a 99% chance of a significant result in a replication study, although this is patently not the case (Oakes, 1986; cf. Haller & Kraus, 2002). More recently, others have documented additional confusions regarding what is to be expected from replications (e.g., Cumming, Williams, & Fidler, 2004).

Because of the importance of replication probability and the confusion surrounding it, recent articles in numerous disciplines have urged researchers to consider replication probability more carefully (e.g., Cumming, 2008; Cumming & Maillardet, 2006; Gorroochurn, Hodge, Heiman, Durner, & Greenberg, 2007; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Killeen, 2005; Robinson & Levin, 1997; Sohn, 1998). Researchers have been offered formulas with which to compute the probability of replicating their current results, and they have been advised to report the resulting replication probabilities as well as—or even in preference to—more traditional statistical measures (e.g., Greenwald et al., 1996; Killeen, 2005; *Psychological Science* editorial board, 2005).

In this article, I consider further the questions of what replication probability is and what factors determine it, and I argue for two main theses. One thesis is that there are two quite different meanings of the term "replication probability," each of which might be of interest to researchers under some circumstances. It is important to be clear about which meaning is under consideration, how-

J. Miller, miller@psy.otago.ac.nz

ever, when discussing replication probability or trying to estimate it, because confusion between the two types of replication probability can lead to inappropriate conclusions. The other thesis is that in practice, neither of these replication probabilities can be estimated at all accurately from the data of an initial experiment, so they are both essentially unknowable. Moreover, the latter thesis implies that researchers are generally ill-advised to summarize their data in terms of estimated replication probabilities, despite the importance of these quantities, because the estimates that they obtain are nearly meaningless.

This article begins with a short review of the standard hypothesis-testing framework in which the question of replication probability often arises. The following sections examine in detail the two different meanings of "replication probability," how each of these probabilities might be estimated, and why the estimates are not very accurate. The General Discussion then considers how the same conceptual distinctions and estimation uncertainty extend to the concept of replication probability within other inferential approaches (e.g., Bayesian).

HYPOTHESIS-TESTING BACKGROUND

Although the framework of null-hypothesis significance testing (NHST) remains controversial (see, e.g., Abelson, 1997; Cohen, 1994; Kline, 2004; Loftus, 1996; Lykken, 1991; Oakes, 1986; Wagenmakers, 2007), even its critics acknowledge that it is still in common use and that many of its problems stem more from misunderstanding and misuse than from inherent flaws. Therefore, replication probability is discussed here mainly within this hypothesis-testing framework. Importantly, this article should not be seen as arguing that NHST is superior to alternative statistical techniques (e.g., confidence intervals; cf. Cumming & Finch, 2005), although I do believe that NHST is one of a wide range of techniques that can usefully be employed, as long as its strengths and limitations are clearly understood.

Within the hypothesis-testing framework, researchers test for a significant effect by computing the probability, under the null hypothesis, of observing data at least as discrepant from the predictions of the null hypothesis as the data they have actually observed. They reject the null hypothesis if this computed probability—sometimes called the "attained significance level" or "p value"—is less than a predetermined cutoff alpha (α) level usually chosen to be $\alpha = .05$. Although some details of this procedure and its rationale differ slightly between the Fisher and Neyman–Pearson schools of inference (see, e.g., Batanero, 2000; Huberty & Pike, 1999), these common features characterize the behavior of practicing researchers, and the differences are not important for the present purposes (cf. Wainer & Robinson, 2003).

As should be known to all who use hypothesis tests, the probability of rejecting a true null hypothesis (i.e., of obtaining a "statistically significant" result by chance) is called the "Type I error probability" or "α level." Correspondingly, the probability of a correct decision to retain a true null hypothesis is $1-\alpha$. The hypothesis-testing procedure is chosen so that the Type I error probability has a certain predetermined value—typically set at $\alpha = .05$, as already mentioned—when the null hypothesis is true.

When the null hypothesis is really false and some alternative hypothesis is true, the probability of rejecting the null hypothesis is called the "power" of the experiment, and the symbol for this probability is $1-\beta$. Correspondingly, under a particular alternative hypothesis, the probability that a false null hypothesis is incorrectly retained is β. As is well known (for a review, see, e.g., Cohen, 1992), power increases with the true size of the effect under study.[1] It also increases with the sample size of the experiment and with the α level associated with the hypothesis-testing procedure. Although the sample size and α level of a given experiment can be specified exactly, the true effect size is never known exactly in practice, precluding direct computation of power.

Researchers generally regard an effect as having been replicated successfully if the effect is statistically significant in both an initial study and a follow-up study, with the results of both studies in the same direction (e.g., larger mean for group A than for group B; Rosenthal, 1993).[2]

TWO MEANINGS OF "REPLICATION PROBABILITY"

It is useful to distinguish between two legitimate but quite different meanings of "replication probability" that might be of interest to researchers under different circumstances. Both may be defined within a frequentist framework. One, which I call the "aggregate" replication probability, is the probability that researchers who obtain significant results in their initial experiments will also obtain significant effects in identical follow-up experiments.[3] As will be discussed in detail, this meaning of replication probability applies across a large pool of researchers working within a common experimental or theoretical context but testing different null hypotheses. It refers to the proportion of successful replications across all of the different null hypotheses tested. The other meaning, which I call the "individual" replication probability, is the long-run proportion of significant results that would be obtained by a particular researcher in exact replications of that researcher's own initial study. This meaning refers to the proportion of significant results within exact replications of a particular initial study (i.e., testing a single null hypothesis), so it is specific to an individual researcher testing that null hypothesis, independent of other researchers working within the same context. Although these two definitions of replication probability may sound nearly equivalent, they are conceptually different, as is developed in the remainder of this section. They are often numerically different as well, and a given study's aggregate replication probability can be either higher or lower than its individual replication probability.

Figure 1 helps illuminate the distinction between the aggregate and individual replication probabilities using a timeline representing an overall research context in which many researchers are working. On the basis of a working theory, each researcher first randomly chooses one of many

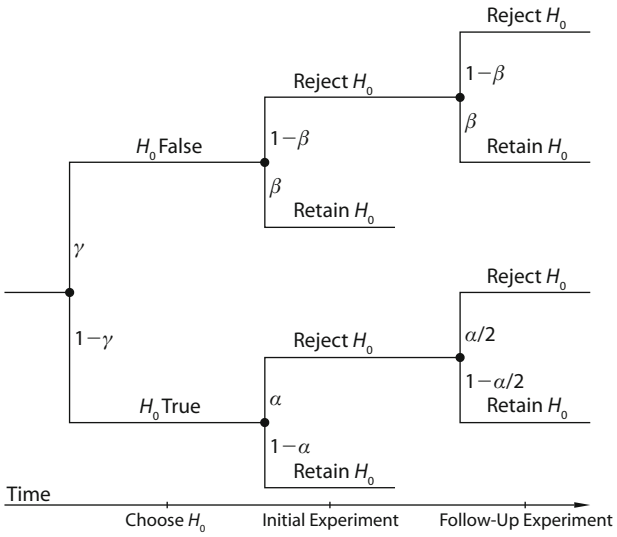

Figure 1. Depiction of the sequence of events within a simple research context. Many researchers carry out experiments within this context, and their experiments are based on a working theory used to generate supposedly false null hypotheses (H_0s). Each researcher first randomly chooses one of these H_0s for test in an initial experiment. With probability γ, the chosen H_0 is indeed false as predicted, and an alternative hypothesis (H_1) is true. With probability $1-\gamma$, the H_0 is actually true (i.e., the theory made an incorrect prediction). When the chosen H_0 is false, as is shown in the top half of the diagram, it is assumed for simplicity that the probability of rejecting H_0 (i.e., experimental power) is always the same, $\Pr(S_1 | H_0 \text{ false}) = 1-\beta$, regardless of the H_0. (This amounts to the assumption that, within this simple research context, H_0 is false to the same degree whenever it is false.) When the chosen null hypothesis is true (bottom half of diagram), the probability of rejecting it is $\Pr(S_1 | H_0 \text{ true}) = \alpha$. Finally, if the initial experiment results in rejection of H_0, the researcher conducts a follow-up experiment to try to replicate the effect. The probability of replication (i.e., of rejecting H_0 in the same direction as in the initial experiment) is again $1-\beta$ for researchers who initially chose a false null hypothesis, and it is $\alpha/2$ for researchers who chose a true null hypothesis.

supposedly false null hypotheses for an experimental test. In the simple model shown in this figure, a randomly selected null hypothesis is false with probability γ and true with probability $1-\gamma$. Once the null hypothesis has been chosen, the researcher conducts an initial experiment to test it. If the null hypothesis is false, the probability of a significant result in the initial experiment is the experimental power, $1-\beta$. If the null hypothesis is true, on the other hand, the probability of a significant result in the initial experiment is α. Finally, if the results of the initial experiment are statistically significant, the researcher carries out a follow-up experiment to see whether the effect is replicated. If the null hypothesis is false, the probability of a significant result in the follow-up experiment is again the experimental power (i.e., $1-\beta$). If the null hypothesis is true, the probability of a significant replication in the follow-up experiment is only $\alpha/2$, because half of the significant-by-chance results will go in the wrong direction (i.e., opposite to the initial result) in the follow-up experiment.

Now consider all of the researchers working in this context who obtain a significant effect in an initial experiment (S_1). What proportion of them will also obtain a significant result in an identical follow-up experiment (S_2)? This aggregate replication probability, p_{ra}, can be computed using standard techniques for working with conditional probabilities (e.g., Krueger, 2001), which are also used to compute the probability that a rejected null hypothesis is actually false (e.g., Ioannidis, 2005). Across all researchers who obtain significant initial results, the aggregate probability of replication is

$$\begin{aligned} p_{ra} &= \Pr(S_2 | S_1) \\ &= \Pr(S_2 \cap S_1) / \Pr(S_1) \\ &= \frac{\Pr(H_1) \cdot \Pr(S_2 \cap S_1 | H_1) + \Pr(H_0) \cdot \Pr(S_2 \cap S_1 | H_0)}{\Pr(H_1) \cdot \Pr(S_1 | H_1) + \Pr(H_0) \cdot \Pr(S_1 | H_0)} \\ &= \frac{\gamma \cdot (1-\beta)^2 + (1-\gamma) \cdot \alpha \cdot \alpha/2}{\gamma \cdot (1-\beta) + (1-\gamma) \cdot \alpha}. \end{aligned} \quad (1)$$

For example, with $\gamma = 0.2$, $1-\beta = .8$, and $\alpha = .05$, the aggregate replication probability (i.e., the conditional probability of a replication in a follow-up experiment, given a significant result in an initial experiment) is $p_{ra} = \Pr(S_2 | S_1) = .645$.

Note, however, that Equation 1 does not yield the long-run proportion of significant replications that will be obtained by any of the individual researchers working in this context, so it does not describe the individual replication probability. For researchers who chose a false null hypothesis, the long-run probability of a replication (i.e., individual replication probability) is simply the experimental power (i.e., $p_{ri} = 1-\beta$). For researchers who chose a true null hypothesis, this probability is half of the Type I error rate (i.e., $p_{ri} = \alpha/2$). In the example of the previous paragraph, then, $p_{ri} = .8$ for some researchers and $p_{ri} = .025$ for other researchers, but it does not equal $p_{ra} = \Pr(S_2 | S_1) = .645$ for any of them. Thus, there are two different values of individual replication probability under the scenario shown in Figure 1, and neither of these equals the aggregate replication probability across all researchers.

To make this distinction more concretely, consider again the example of $\gamma = 0.2$, $1-\beta = .8$, and $\alpha = .05$. Of 1,000 researchers working within this context, 200 conduct experiments in which H_0 is false, and $.8 \times 200 = 160$ of these obtain significant results (ignoring binomial variability). For each of these 160 researchers, the individual replication probability is .8—namely, the power associated with their experiments—because this is the long-run probability of getting significant results in identical follow-up experiments. Thus, 128 of the 160 should successfully replicate their findings. The other 800 researchers conduct experiments in which the null hypothesis is true, and only $.05 \times 800 = 40$ of these obtain significant results. For these 40 researchers who made Type I errors, the individual replication probability is $\alpha/2 = .025$, because half of the significant results will go in the wrong direction, so only one of the researchers should replicate the initial result. The aggregate replication probability of $p_{ra} = \Pr(S_2 | S_1) = .645$ is the probability of a significant result in a follow-up experiment that is randomly selected

from among the experiments with significant initial results [i.e., (128 + 1) / (160 + 40) = .645]. Note that this probability is also the weighted average of the individual replication probabilities across the 200 researchers who obtained significant results in the initial experiment (.8 × 160 + .025 × 40) / 200. Now, a given researcher in this scenario would have no way of knowing whether an obtained significant effect was real or a Type I error, and might therefore decide to regard this aggregate replication probability as an estimate of that effect's individual replication probability. Nonetheless, it should be kept in mind that this value actually reflects the probability of a significant effect across replications of many different experiments testing different null hypotheses, not across many different replications of a single experiment testing the same null hypothesis that was tested initially.

Another way to illuminate the distinction between individual and aggregate replication probabilities is to consider the probability of obtaining $j \geq 2$ successful replications. If the probability of single replication is p_1, one might expect the probability of j independent replications to be p_1^j. This expectation is correct for individual but not for aggregate replication probabilities. For a researcher whose individual replication probability is $1-\beta = .8$, for example, the probability of j replications is $.8^j$, because the replications are all independent realizations of that researcher's particular experiment, each of which has the same power. The same formula does not apply for aggregate replication probability, however, because multiple replications of a particular experiment are dependent, in that they all test the same null hypothesis. For a concrete illustration of the aggregate probability of j replications, consider further the 1,000 researchers working within the scenario illustrated in Figure 1, again with $\gamma = 0.2$, $1-\beta = .8$, and $\alpha = .05$. Of the 128 researchers discussed previously who tested a false H_0 and then successfully replicated their findings in a follow-up experiment, .8 × 128 = 102.4 will also be successful in a second replication attempt. Of the 1 researcher who tested a true H_0 and successfully replicated the findings in a follow-up experiment, .025 × 1 = .025 will also be successful in a second replication attempt. Thus, the aggregate probability of two successful replications, given a significant initial result, is (102.4 + 0.025) / (160 + 40) = .512. This probability is much greater than the square of the aggregate replication probability for a single replication (i.e., $.512 > .645^2 = .416$), illustrating that the aggregate probability of two replications is not p_{ra}^2. See Appendix A for further information and illustrations concerning the dependence of aggregate probability on the number of replications, j.

In summary, "replication probability" can be used in either of two senses. The *aggregate* replication probability is the probability of a significant result when replicating a randomly selected effect out of a large pool of different significant effects, whereas the *individual* replication probability is the probability of a significant result across many identical attempts to replicate a single significant effect. The aggregate replication probability depends on the larger research context, including all of the effects in the pool of initially significant ones, whereas the individual replication probability depends only on the one particular effect that is under consideration.

These two senses of "replication probability" are relevant for answering different questions. A researcher who considers replicating a previously observed effect probably wants to know about the long-run probability of replicating that particular effect—that is, its individual replication probability. As I will discuss in the Estimation of Individual Replication Probability section, techniques have been suggested for summarizing the data from an initial experiment to estimate such individual replication probabilities. Most of these techniques simply ignore the idea that the experiment under consideration was selected out of some larger pool (e.g., the scenario depicted in Figure 1). In contrast, the concept of an aggregate replication probability is relevant for a researcher asking how likely it is that significant results in a particular research area actually represent spurious findings or Type I errors (e.g., Oakes, 1986, Table 1.2.1). In this case, the question involves the whole research area, and it must be answered by considering the aggregate proportion of significant results obtained from real effects versus Type I errors within that area. If more of the experiments in the area test true null hypotheses, then—as will be discussed in the Aggregate Replication Probability section—it is less likely that a significant initial effect will be replicated.

Note that the individual replication probability associated with a particular experimental design is exactly the same as the probability of a significant result (in the observed direction) in the initial experiment, because it is simply the probability of rejecting the null hypothesis (in this direction) in an experiment of this type. For example, if the probability of a significant result in an initial experiment with this design was .5, then the probability is also .5 for all identical follow-up experiments. One simple justification for this claim is that the initial and follow-up experiments are independent samples from the pool of all possible replications of that experiment, so their results do not depend on the order in which they are carried out. A priori, the follow-up experiment is just as likely to have a smaller p value than the initial experiment as it is to have a larger one.[4] If the null hypothesis is false, the individual replication probability is simply the experimental power. If the null hypothesis is true and a two-tailed test is used, this probability is $\alpha/2$ (i.e., typically .025). With a true null hypothesis, a significant effect will be observed with probability α, but half of the significant results will go in the wrong direction, as noted earlier.

It may be somewhat counterintuitive that the individual replication probability is the same as the probability of getting a significant result in the initial experiment, because psychologically they seem quite different. A researcher might reason, for example: "Before I ran my initial experiment, I would have said it was only 50/50 that I would get this effect. Now that I have run the experiment and gotten the effect at $p < .01$, surely my odds of getting it again have been improved by these results!"

This reasoning is perfectly valid if the researcher is considering the "odds of getting it again" in terms of the aggregate replication probability. The initial significant

result suggests that the effect under study is more likely to belong to some population of real effects than to some other population of spurious ones, and of course aggregate replication probability is higher in the former population than in the latter. At the same time, however, the reasoning is invalid if it is applied to the individual replication probability. The individual replication probability—that is, the long-run probability of a significant result with a given experimental design—simply does not change when that experiment is repeated. The results of the initial experiment may reveal something about the value of power, but they do not change that value. This point can be illustrated by an analogy: Suppose we select one coin from a large set of coins—some of which may be biased—flip the selected coin 50 times, and obtain 30 heads. Under these conditions, what is the probability of heads on toss number 51? In an aggregate sense, the results of the first 50 tosses inform us that this coin is more likely to have come from a population of coins with a bias toward heads. Clearly, though, the probability (i.e., long-term frequency) of this coin coming up heads on the next toss is the same as it was in the first 50 tosses. The observation of 30 heads provides some information about what that probability is (and has been all along), but it does not change that probability. The point is that the results of the initial experiment only inform us about the power and individual replication probability—they do not change it.

It is worth being explicit that individual replication probability remains constant, because people seem intuitively prone to regard many kinds of probabilities—including replication probabilities—as fluctuating depending on prior outcomes. True probabilities are usually unknown, so we estimate them, and of course our estimates do change on the basis of new information, even though the probabilities themselves do not. The "hot hand" and gambler's fallacies provide two obvious examples of how tempting it is to believe that probabilities can change across repeated events (see, e.g., Ayton & Fischer, 2004; Boynton, 2003; Sundali & Croson, 2006). Worse yet, statistical formulas and terminology sometimes encourage this misconception. For example, Laplace's "law of succession" says that "if we have experienced S successes and F failures out of $S + F = n$ trials, the chance of success on the $(n+1)$st trial is $(S + 1)/(n + 2)$" (Wilson, 1927, p. 210).[5] Taken literally, this "law" certainly suggests that the probability of success changes with each new trial, although really the law describes fluctuations in the best estimate of this probability rather than in its true value. The temptation to think of individual replication probabilities as changing may be especially strong because the aggregate replication probability does change with each new experimental outcome. Aggregate replication probability changes, though, because each new outcome changes the pool within which the aggregate probability is computed, not because the outcome changes the individual probability for any particular effect under study.

An important corollary of the idea that individual replication probability remains constant is that—although we may be able to estimate individual replication probability from the initial data, as is considered in the next section—it is impossible to calculate the true individual replication probability without making quite specific assumptions. One way to see this is to note that any given observed set of data might have been obtained under many different states of the world, so these data by themselves can never specify exactly which state of the world gave rise to them. Each different state of the world corresponds to a different individual replication probability, so the data simply do not uniquely determine the exact probability of replicating the results. To continue with the coin-flipping analogy, for example, we might easily have gotten 30 heads in 50 tosses from a coin with any true P within the range of at least (say) .58–.62. Since the exact individual replication probability depends on the exact P, we cannot recover it from the observed results.

ESTIMATION OF INDIVIDUAL REPLICATION PROBABILITY

Writing about the poor odds of replicating a particular significant result, Rosenthal (1993) said "A related error often found in the behavioral and social sciences is the implicit assumption that if an effect is 'real,' we should therefore expect it to be found significant again on replication. Nothing could be further from the truth" (p. 542). To illustrate that point, he considered an example of a researcher working at a power level of .5, noting among other dismal facts that for this case "there is only one chance in four that both the original investigator and a replicator will obtain [significant results]" (p. 543).

Rosenthal's (1993) comments, though of course entirely correct as stated, seem to imply that individual replication probability can be estimated quite precisely, and that it is rather low. Note, however, that his assessment of power (and hence of individual replication probability) made no reference to the data of the initial experiment. Instead, he simply assumed that the initial experiment had a power level of .5, in which case the follow-up would, too.

How, then, can one use the data of the initial experiment to estimate the individual replication probability with the same experimental design? As was noted earlier, experimental power depends on the sample size and α level, which are known, and on the effect size, which is unknown. The standard approach is to assume that the true effect size is equal to the effect that was actually observed in the initial experiment (e.g., Gorroochurn et al., 2007; Greenwald et al., 1996; Oakes, 1986; Rosenthal, 1993). From this assumed true effect size, it is straightforward to compute power as a function of sample size and α level (see, e.g., Cohen, 1988, for relevant formulas, or Faul, Erdfelder, Lang, & Buchner, 2007, for a computer program that performs such computations). Estimates computed using this approach are sometimes referred to as "post hoc" (e.g., Onwuegbuzie & Leech, 2004) or "observed" power values (SPSS, 2006).

For example, consider tossing a coin 100 times to test the null hypothesis that the true probability of heads is .5. Binomial tables indicate that this null hypothesis can be rejected ($p < .05$, two-tailed) if the observed number of heads is greater than 60 or less than 40. Suppose that 65

heads are obtained, a value that is sufficient to reject the null at $p = .004$, two-tailed. To estimate power, one then assumes that the true probability is $P = .65$ (i.e., the observed value). With that true P, binomial tables indicate that the probability of obtaining more than 60 heads is .83, so this is the estimated power of 100-toss experiments with this coin—both the initial experiment and all follow-ups.

In the absence of any other information about the true effect size, it may seem reasonable to estimate power—and hence individual replication probability—by assuming that the true effect equals the effect observed in the initial experiment (e.g., Posavac, 2002), although problems with this approach have sometimes been stressed within the statistical literature (e.g., Hoenig & Heisey, 2001). Critically, as with any value estimated from data, the resulting value is only an estimate of power, not the true value (e.g., Froman & Shneyderman, 2004; Macdonald, 2003; Sohn, 1998). The observed effect size is subject to sampling error, so it is rarely exactly equal to the true effect size. Consequently, the estimated power is unlikely to equal the true power. Instead, the true power will be less than the estimated power if the observed effect is larger than the true effect, and the true power will be more than the estimated power if the observed effect is smaller than the true effect. Inaccurate power estimates are a direct consequence of variability in the observed effect size, so they affect not only this simple power estimate, but other, more sophisticated ones as well (e.g., Cumming, 2008, Appendix B). For example, as a consequence of this variability, two identical replications of the same experiment will produce different data, and thus different estimates of power, even though the true power is the same for both (according to the definition of "identical replications"; cf. Hoenig & Heisey, 2001). Conversely, two instances of different experiments—having true power levels that are actually quite different—may produce identical observed effects, and thus yield identical estimates of power.

How different might the estimated and true individual replication probabilities (i.e., power levels) be? Because the only random quantity influencing the power estimate is the observed effect size, this question can be answered by looking at a 95% confidence interval for the effect (Froman & Shneyderman, 2004). Power increases with effect size, so an upper bound for power can be estimated by assuming that the true effect is the largest value in the confidence interval. Similarly, a lower bound for power can be estimated by assuming that the true effect is the smallest value in the interval. As is illustrated by the following example, these estimated bounds for power will capture its true value 95% of the time, just as the confidence interval for the effect captures its true size 95% of the time (e.g., Cumming & Finch, 2001).

For example, consider again tossing a coin 100 times and obtaining 65 heads, a scenario depicted in Figure 2. As was discussed earlier, .65 is the point estimate of P, and that value corresponds to an estimated individual replication probability of .83. A standard 95% confidence interval for P, however, indicates that the true P may be as small as .56 or as large as .74 (see Appendix B for details). If the true value of P is really only .56, the individual replication probability is really only .18. This is a worst-case estimate of individual replication probability, because it is computed using the estimated P value closest to that specified by the null hypothesis. At the other extreme, if the true value of P is .74—the largest value in the confidence interval—then the individual replication probability is .998. This is correspondingly the best-case estimate of individual replication probability, computed with the estimated P farthest from the value specified by the null hypothesis. Critically, any experiment's true individual replication probability will fall between the estimated upper and lower power bounds on the vertical axis if and only if its true P value falls between the upper and lower bounds on the horizontal axis, and vice versa, as is evident from the geometry of Figure 2. Given that a confidence interval for P captures the true P value in 95% of all experiments (e.g., Cumming & Finch, 2001), it follows that the estimated upper and lower bounds for individual replication probability will also capture the true replication probability in 95% of all experiments.

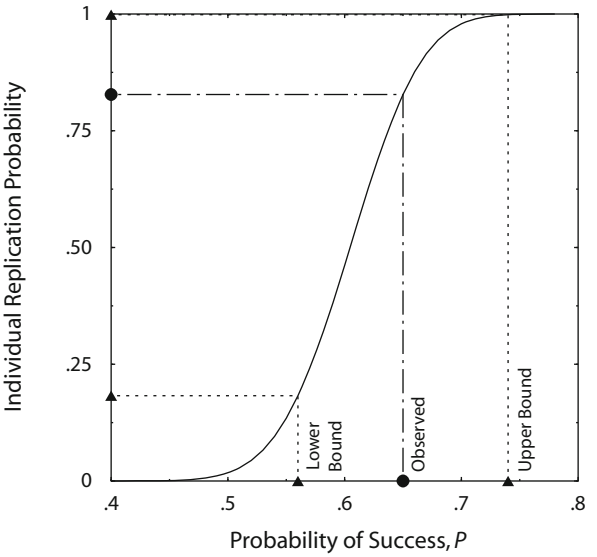

Figure 2. Illustration of estimated upper and lower bounds for individual replication probability in a binomial experiment with 65 successes in 100 trials. The solid ogive shows the individual replication probability (i.e., the probability of 61 or more successes) as a function of the true probability of success in each trial. As indicated by the circles, the actual experiment yielded .65 as the observed proportion of successes; assuming the true *P* is .65 yields an estimated individual replication probability of .83. With an observed proportion of .65, however, the 95% confidence interval for the true *P* extends from the lower bound of .56 to the upper bound of .74, as indicated by the triangles. Corresponding lower and upper bounds for the individual replication probability are .18 and .998. Because individual replication probability is monotonically related to the true *P*, the true individual replication probability falls between its estimated bounds if and only if the true *P* falls between its estimated bounds.

In summary, the overall conclusion from a confidence-interval-based analysis of this binomial example is that the individual replication probability, initially estimated to be .83, could actually be as low as .18 or as high as .998—quite a wide range. Even though it is reasonable to

expect the individual replication probability to be nearer to .83 than to .18 or .998, the possibility that it could be anywhere in this wide range means that we should not place very much faith in the original point estimate of .83. It seems that the initial result of 65 successes actually reveals hardly anything about the true individual replication probability. In other words, this initial result is consistent with a wide enough range of true *P* values that the individual replication probability could be almost anything.

Computations analogous to those illustrated with the preceding binomial test were carried out for a variety of sample sizes and observed numbers of successes, and Figure 3A summarizes the results. Computations were carried out for $N = 100$, 500, or 1,000 trials and for every possible statistically significant result at each sample size. Each significant result was plotted on the horizontal axis in terms of its *p* value (e.g., obtaining 65 successes in 100 trials represents a *p* value of .004). The three different estimated individual replication probabilities associated with each statistically significant outcome were plotted on the vertical axis, with these probabilities estimated from the observed proportion or from the upper or lower bound of the 95% confidence interval for the true proportion. With 65 successes in 100 trials, for example, the three estimated replication probabilities already noted are the values of .998 (solid line), .83 (dashed line), and .18 (dot-dashed line), indicated by the three arrows. Similarly, other points along the three $N = 100$ curves represent outcomes with 61–81 successes, and points along the $N = 500$ and $N = 1{,}000$ curves represent observed numbers of successes corresponding to the indicated *p* values with those sample sizes. All three estimated replication probabilities are almost completely determined by the *p* value independently of the sample size, so the curves for $N = 100$, $N = 500$, and $N = 1{,}000$ overlap almost perfectly.

The most remarkable feature of the results shown in Figure 3A is the wide range of individual replication probabilities that are consistent with a given statistically significant experimental result. For experiments yielding *p* values in the range of approximately .005 to .05, the individual replication probabilities associated with the upper and lower bounds for the true *P* value cover almost the entire 0–1 range. In these cases, then, the significant result of the initial experiment actually provides almost no constraint on the true individual replication probability. The individual replication probability is tightly constrained only by very highly significant initial results, which yield upper and lower bounds near 1.0, at the upper left of each panel. In short, individual replication probability is known fairly precisely only when the effect is so large that this probability is near 1. Moreover, this pattern seems to hold virtually independently of sample size. Although one might expect higher replicability for larger samples, this expected effect is absent from the figure because of a trade-off between sample size and effect size. Specifically, to hold the *p* value constant, the observed effect must be made smaller as the sample size becomes larger. Using smaller effects for larger samples overcomes the replicability advantage that would otherwise be expected with the larger samples (cf. Cumming, 2008).

To show that these results are not associated with some peculiarity of the binomial test, Figures 3B and 3C show analogous computations for hypothetical experiments that would be analyzed using two other statistical tests. Figure 3B shows computations for experiments in which a *t* test would be used to test the null hypothesis that a true mean or a true difference between means, Δ, equals zero. Figure 3C shows computations for experiments using an observed sample correlation coefficient to test the null hypothesis of a zero true correlation in the population as a whole (i.e., H_0: $\rho = 0$). Again, for each possible significant observed *t* value (panel B) or significant sample correlation *r* (panel C), the *p* value was computed and used to determine the location on the horizontal axis. Three individual replication probability estimates were computed for each result, assuming that the true effect was equal either to the observed value (*t* or *r*) or to the upper or lower bound of a 95% confidence interval for the true effect (see Appendix B for details).

The results shown in Figures 3B and 3C are virtually identical to those in Figure 3A. Again, results significant at the level of .005–.05 are consistent with effect sizes for which individual replication probability could be anywhere in the range of approximately .1–1, with little effect of the sample size parameter as it varies within each panel. Thus, the conclusions from the binomial test seem to generalize perfectly well to *t* tests and tests of correlation.

The overall conclusion reached by looking at confidence intervals for observed effects is that only the most highly significant results of an initial experiment really provide any useful information about individual replication probability (or, equivalently, about post hoc or observed power), regardless of the sample size or the type of statistical test. Although researchers can estimate individual replication probability by assuming that the true effect matches the observed one, in practical terms the error associated with the observed effect is usually so large that such estimation seems pointless. Researchers need to be wary, then, of precise statements about individual replication probability, such as "After obtaining $p = .03$, there is actually only a 56.1% chance that a replication will be statistically significant with two-tailed $p < .05$" (Cumming, 2008, p. 287) and "a *p* value of .005 (note the extra zero) means the probability of exact replication is .80" (Harris, 1997, p. 10; for similar claims, see, e.g., Gorroochurn et al., 2007, p. 327, and Greenwald et al., 1996, p. 181). As was noted by Froman and Shneyderman (2004), despite recent calls for more emphasis on estimating power from available data (e.g., Onwuegbuzie & Leech, 2004), the same caveat also applies to the post hoc power estimates provided by newer statistical software packages. Indeed, even qualitative statements such as "replicability is closely related to the *p* value of an initial study" (Greenwald et al., 1996, p. 180) must be regarded as broad generalizations with little diagnostic value for any specific experimental result. To get a reasonably accurate estimate of individual replication probability requires much tighter bounds on the true effect size than are usually provided by statistically significant results.

Another remarkable feature of the estimated individual replication probabilities that is difficult to see in Figure 3

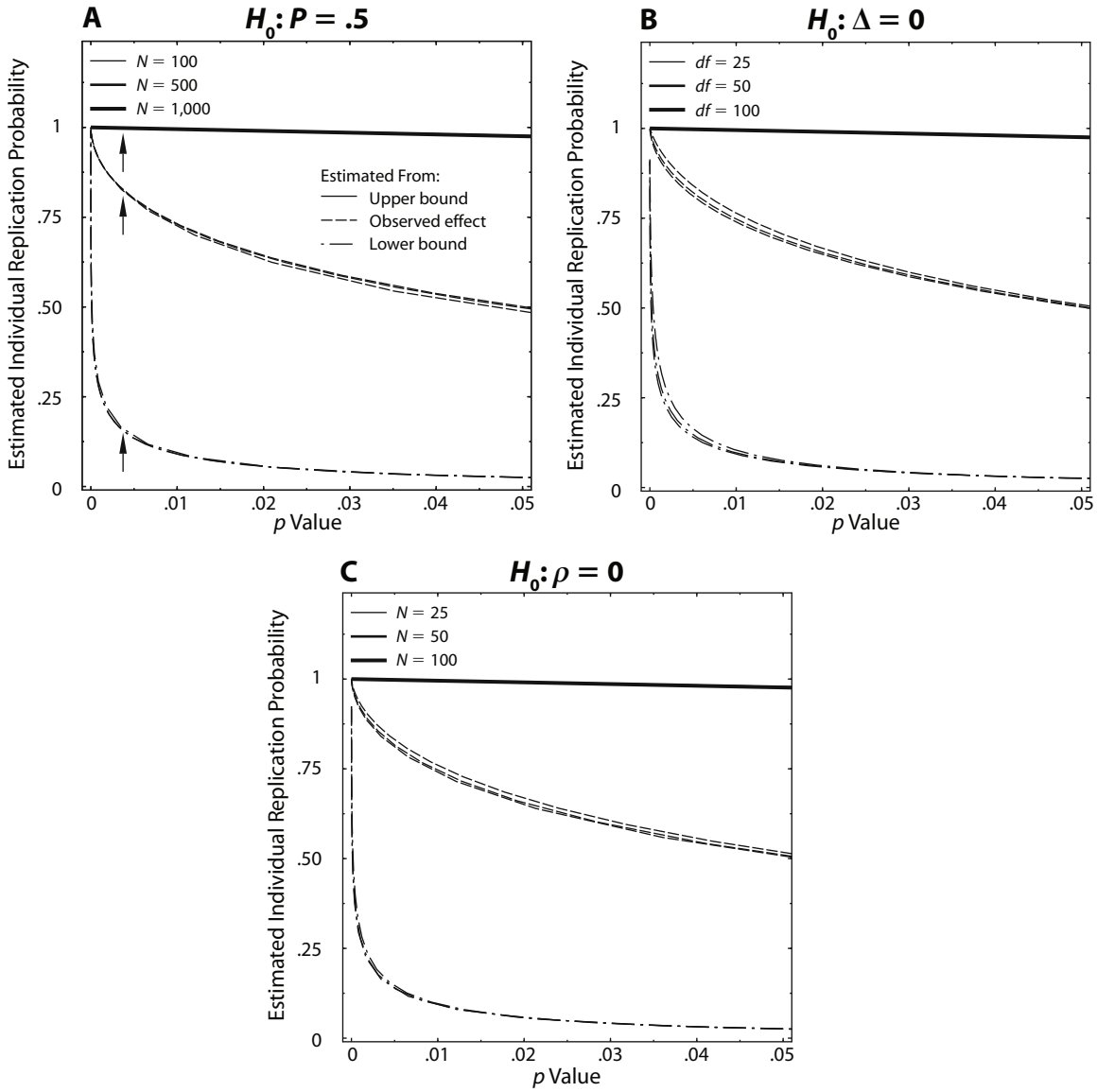

Figure 3. Three estimates of the probability of rejecting the null hypothesis in an identical replication experiment as a function of the *p* value of the initial experiment. The solid lines at the top of each panel show individual replication probabilities estimated by assuming that the true effect is at the upper bound of the 95% confidence interval for the true effect; these are the best-case replication probabilities. The dashed lines in the middle show probabilities estimated by assuming that the true effect exactly matches the effect observed in the initial experiment. These are standard point estimates of individual replication probability. The dot-dashed lines at the bottom show individual replication probabilities estimated by assuming that the true effect is at the lower bound of the confidence interval for the true effect; these are the worst-case individual replication probabilities. (A) Estimated probability of rejecting the null hypothesis $P = .5$ using a binomial test with sample sizes of $N = 100$, 500, and 1,000. The three points marked with arrows indicate the values corresponding to the example of 65 successes in 100 trials, as discussed in the text. (B) Estimated probability of rejecting the null hypothesis that a mean or difference of means $\Delta = 0$ using a *t* test with 25, 50, or 100 degrees of freedom (*df*) for error. (C) Estimated probability of rejecting the null hypothesis that a true correlation $\rho = 0$ for sample sizes of $N = 25$, 50, or 100.

is that the worst-case individual replication probability is only slightly greater than .025 when the *p* value is .05. It seems astonishing that individual replication probability could actually be less than the Type I error rate—even after getting a significant result in an initial experiment—but exactly the same pattern was obtained for all sample sizes with all three statistical tests. In retrospect, it is not difficult to see why this happens with two-tailed tests. When an observed result is just significant at $p = .05$, the true effect could be infinitesimal, because the lower bound of the confidence interval for the effect is only slightly different from the true value specified by the null hypothesis. Therefore, the probability of a significant result (in the same direction) is only slightly higher under this assumed worst-case bound than it is under the null hypothesis (Sohn, 1998). For a two-tailed test with $p = .05$, the probability assigned

to that tail under the null hypothesis is .05/2 = .025, so the probability associated with the same tail is only slightly larger than this when the true value is only slightly different from that specified by the null hypothesis.

AGGREGATE REPLICATION PROBABILITY

The previous section considered estimating the individual replication probability from a significant initial experiment considered in isolation. Real experiments are conducted within a larger research context, however—not in isolation—and in some circumstances it may be desirable to estimate the aggregate probability of replicating a randomly selected experimental effect within this context (e.g., for the purposes of a literature review).

For example, I claimed earlier that when a single experiment is considered in isolation, the *p* value of the replication is just as likely to be smaller than the initial result as it is to be larger, because the order of the two experiments has no bearing on their relative *p* values. A skeptic might reply that this claim has no relevance when evaluating research within a given area, because the initial results being considered for possible replication have already been selected for being significant. Such selection obviously introduces a bias in favor of significant initial results. Therefore, the skeptic might argue, the replication is likely to yield a larger *p* value than the initial experiment. The force of this argument and the strength of its effect (i.e., how much larger the replication *p* value is likely to be) depend critically on aspects of the research context.

In this section, I examine some aspects of the research context that have implications for aggregate replication probability. Given any specific set of assumptions about a particular research context, aggregate replication probability can be computed using a Bayesian approach, as was illustrated with Equation 1. Unfortunately, in practice, researchers virtually never have the information about the research context required to perform such computations. Nonetheless, it is illuminating to see how strongly aggregate replication probability would depend on such information if it were available. The strong dependence of aggregate replication probability on unavailable information shows that aggregate replication probabilities are generally unknowable, just like individual replication probabilities.

Theory Strength and Aggregate Replication Probability

The theoretical basis of the initial study is one aspect of the research context with major implications for aggregate replication probability. A researcher's choice of experiments is always guided by some theory, whether formal or intuitive. Typically, a researcher's theory suggests that a large number of effects should be real, and the researcher selects one of those suggested effects for experimental test. The researcher then conducts an initial experiment to test for the selected effect, and—if a significant effect is obtained—conducts a follow-up experiment (cf. Figure 1). For convenience in modeling this process within a frequentist framework, I will pretend that the researcher randomly selects the effect to test in the initial experiment from among the set of all effects suggested by the guiding theory.

It is easy to see that aggregate replication probability depends markedly on the quality of the theory that led to the initial experiment. As an extreme example, researcher W might work with a theory so weak that none of its suggested effects are in fact real (i.e., the null hypothesis is true in all cases, corresponding to $\gamma = 0$ in Figure 1). All of researcher W's significant results would be Type I errors, and the probability of replicating any one of them (in the same direction) would always be just $\alpha/2$. Thus, for this researcher the skeptic is quite right: The initial data would have produced an effect of $p \leq .05$ only by chance, so the replication's *p* value would very likely be larger.

At the other extreme, researcher S might use a theory so strong that all of its suggested effects are real (i.e., $\gamma = 1$ in Figure 1). Assume further that these real effects are so large that power is $1-\beta = .9$ for a typical experiment. All of researcher S's significant results would be correct rejections, and the probability of replicating any one of them (in the same direction) would be .9. Clearly, aggregate replication probability would be much higher for researcher S than for researcher W, regardless of the *p* values of their initial experiments. Of course, researcher S would be far more likely to get significant results in the initial experiment than researcher W, but even researcher W would get some significant findings occasionally. In fact, researcher S would usually get results that were significant at *p* values well below .05, because experiments with such high power tend to produce quite low *p* values. With a *t* test, for example, experimental setups having a power of .9 typically yield median two-tailed *p* values of approximately .001–.005. For researcher S's initial result that was significant at $p = .05$, then, the skeptic would be quite wrong; this researcher's replication would usually achieve a lower *p* value than the initial $p = .05$.

Presumably most researchers work with theories intermediate in strength between these weak and strong extremes. Critically, though, a very specific assumption about theory strength is always needed to compute aggregate replication probability. This point is explicitly recognized within Bayesian evidence accumulation approaches, in which the Bayesian "prior" distribution for the true effect size is exactly this sort of assumption. Bayesian analyses of scientific evidence accumulation are usually developed as an alternative to NHST rather than as a complement to it, however (e.g., Falk, 1998; Killeen, 2006; Wagenmakers, 2007), so the consequences of theory strength for NHST have rarely been considered (Krueger, 2001, Figure 2; Macdonald, 2005). It may therefore be helpful to consider some examples illustrating the influence of this factor on values of aggregate replication probability within NHST. Ioannidis (2005) presented similar examples illustrating the influence of the same factor on the probability that a rejected null hypothesis is actually false.

To get an idea of the quantitative influence of theory strength on aggregate replication probability, consider two researchers conducting experiments with $N = 100$ binomial trials to test the null hypothesis that $P = .5$ within

their own separate research contexts. Suppose that 90% of the effects suggested by researcher S's theory are real (i.e., $\gamma = .9$), whereas only 10% of the effects suggested by researcher W's theory are real (i.e., $\gamma = .1$). For both researchers, assume that $P = .60$ for a real effect; binomial tables indicate that this true P value yields a power level of $1-\beta = .4621$. Now suppose that each researcher conducts an experiment and observes 63 successes; what is the probability that each will successfully replicate the results in a follow-up experiment?

Using Bayes's theorem and binomial tables (for further details, see Appendix A), it is possible to compute the conditional probability that a randomly sampled null hypothesis is true, given this experimental outcome within this scenario. For researcher S, the result is

$$\Pr(H_0 \text{ true} \mid 63 \text{ successes}) = .0044.$$

For this researcher, the aggregate replication probability can then be computed as

$$\Pr(\text{replication} \mid 63 \text{ successes}) = .4602.$$

Parallel computations show that the aggregate replication probability is only .3473 for researcher W, given the same observed result of 63 successes. Thus, aggregate replication probability is more than 30% higher for researcher S than for researcher W, despite the fact that their initial results were identical.

Figure 4 summarizes the results of analogous computations under a variety of scenarios. This figure shows aggregate replication probability as a function of (1) the p value in the initial experiment, (2) the probability that the tested effect is actually real (i.e., $\gamma = .1, .3, .5, .7$, or .9), and (3) the size of the true effect when the null hypothesis is false. A "large effect" was assumed to be an effect that, if observed in a sample, would be significant at the .001 level. A "small effect" was assumed to be one that would be significant at the .05 level. These large and small effect sizes correspond to power ($1-\beta$) levels of approximately .9 and .5, respectively. The panels in the top, middle, and bottom rows present computations for three different hypothesis-testing procedures (binomial, t test, and correlation). The panels on the left show computations for experiments with smaller sample sizes, and those on the right show computations for experiments with larger sample sizes. Once again, the patterns are remarkably consistent across sample sizes and hypothesis-testing procedures.

Figure 4 shows that aggregate replication probability depends not only on the p value in the initial experiment but also—and quite strongly—on the strength of the theory used to generate the experimental hypothesis in the first place. For one thing, aggregate replication probability is higher if more of the theory's suggested effects are indeed real, with aggregate replication probability being especially low in the examples with only a $\gamma = .1$ probability that a suggested effect is real. For another, aggregate replication probability is typically much larger when a theory correctly predicts larger effects (solid lines) than when it correctly predicts smaller ones (dashed lines). This is to be expected, of course, since power increases with effect size. Note, however, that aggregate replication probability is larger with larger effects to a much greater degree when the p value is quite small. For p values near .05, in contrast, aggregate replication probability can actually be larger when the effects are small than when they are large. This is because an observed effect of "only" $p = .05$ may be equally consistent, or even more consistent, with the null hypothesis than with the existence of a large effect.

Unfortunately, the effects on aggregate replication probability shown in Figure 4 are not quantitatively useful to researchers in practice, because there is no reason to accept the particular assumptions used in computing these probabilities. Indeed, the idea of a dichotomous effect—present to a particular degree or not at all—is rather implausible, because most theories predict effects that are present to different degrees (i.e., some predicted effects are large and others are small). Instead, the figure's results are important because they demonstrate how much aggregate replication probability can depend on the strength of the underlying theory. Indeed, with the assumed variation in theory strength and with initial p values in the .01–.05 range, aggregate replication probability depends more strongly on theory strength than on the actual experimental results. Given that the strength of the theory is virtually always unknown in practice, this dependence reinforces the view that researchers have very little basis on which to determine aggregate replication probability, despite the availability of appropriate computational formulas (e.g., Equation 1).

Number of Opportunities for a Significant Result and Aggregate Replication Probability

Another important aspect of the research context that influences aggregate replication probability is the number of opportunities for obtaining a significant result. In real experimental work, for example, researchers often make several attempts at rejecting a suspect null hypothesis before obtaining a significant result. The unsuccessful early attempts tend to be regarded as pilot studies whose nonsignificant results provide information used to improve the experimental protocol. When the null hypothesis is actually true, though, these multiple attempts simply provide multiple opportunities for making a Type I error. Another typical situation with multiple opportunities to reject null hypotheses arises when researchers test a number of different null hypotheses within a single experiment (e.g., testing all possible pairwise correlations among three or more measured variables).

As has often been discussed in the literature on hypothesis testing, all scenarios with repeated hypothesis tests provide multiple opportunities for Type I errors (see, e.g., Shaffer, 1995). This fact is important for the present purposes because aggregate replication probability is smaller when there is a higher probability that the initial result was a Type I error. Unfortunately, as will be seen in this section, it is possible to correct for the contaminating effect of number of opportunities (e.g., with a Bonferroni adjustment) only by making strong and generally untestable assumptions about the strength of the theory underlying the initial experiment.

To get an idea of the size of the "number of opportunities" effect, consider researchers working with a theory in

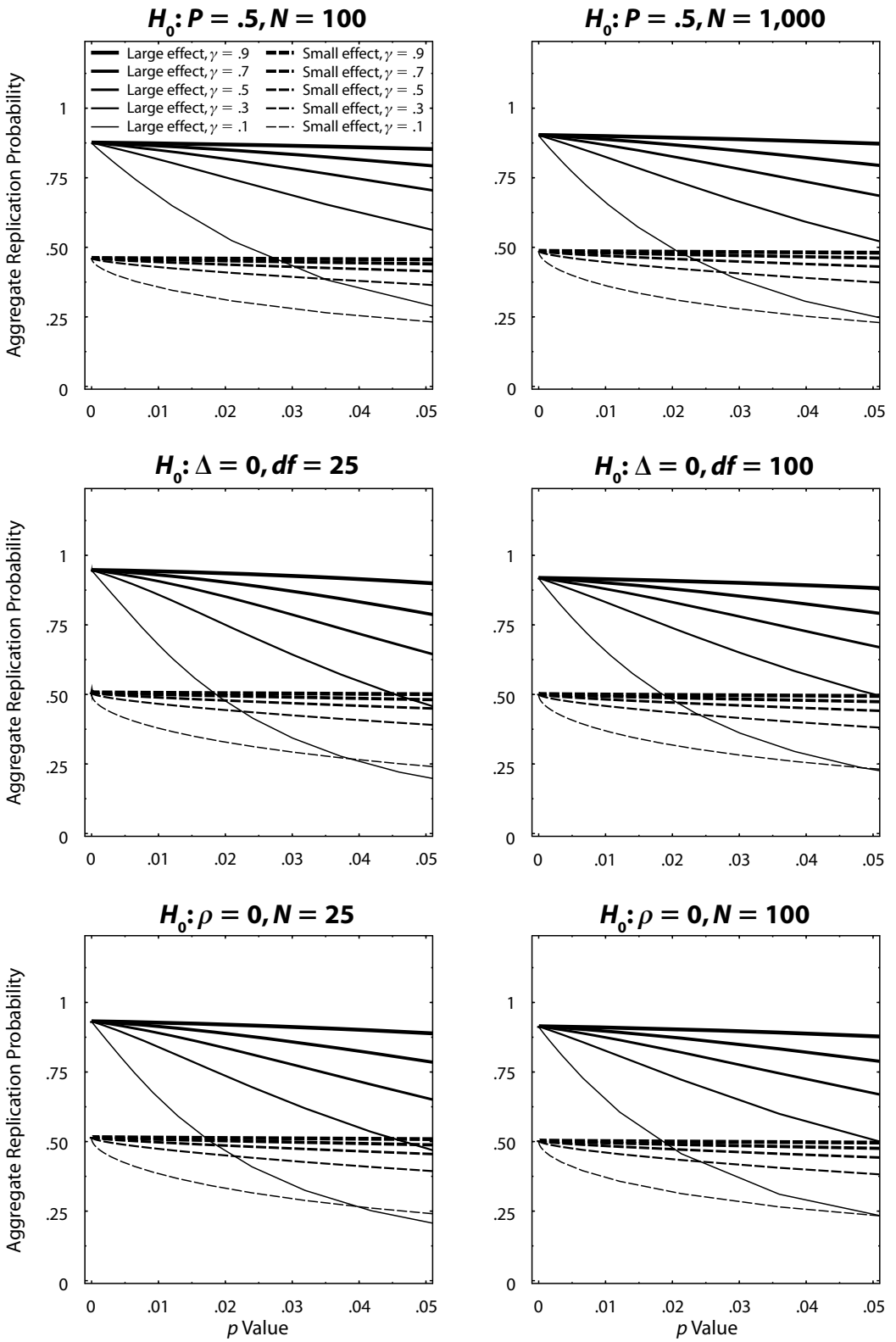

Figure 4. Aggregate replication probability as a function of the *p* value of the initial experiment and the strength of the background theory on which the initial experiment was based. The line thickness represents the probability that a suggested effect is real (γ), varying from $\gamma = .9$ for the thickest lines to $\gamma = .1$ for the thinnest ones. Solid lines represent theories for which the real effects are larger, whereas dashed lines represent theories for which these effects are smaller. (Top) Probability of rejecting the null hypothesis $P = .5$ using a binomial test with the indicated sample size of $N = 100$ or 1,000. (Middle) Probability of rejecting the null hypothesis that a mean or difference of means $\Delta = 0$ using a *t* test with the indicated 25 or 100 degrees of freedom (*df*) for error. (Bottom) Probability of rejecting the null hypothesis that a true correlation $\rho = 0$ for the indicated sample size of $N = 25$ or 100.

which exactly half of the suggested effects are real (i.e., $\gamma = .5$). Suppose that each researcher tests a given null hypothesis across k opportunities and obtains exactly one significant result. For any assumed real effect size, Bayes's theorem can again be used to compute aggregate replication probability, conditional not only on the p value of the initial result but also on the number of opportunities (see Appendix A for further details). Figure 5 displays the results of such computations for the same hypothesis-testing procedures and sample sizes used in constructing Figure 4.

The results in Figure 5 reveal quite a strong decrease in aggregate replication probability as the number of opportunities increases, especially when the p value of the initial result is in the range of .05 to .01. Furthermore, the figure reveals an interaction between the number of opportunities and the size of the theory's effects that may be rather counterintuitive. Aggregate replication probability decreases rapidly with the number of opportunities for theories that predict large effects, whereas it decreases rather slowly for theories that predict small effects. This interaction is so large that aggregate replication probability can actually be lower when theories suggest large effects (solid lines) than when they suggest small effects (dashed lines). This surprising result can be understood by considering the implications of the unsuccessful opportunities (i.e., failures to get significant results). If an effect is large when it is real, several failures actually provide rather strong evidence that there is no real effect and, hence, that the significant effect eventually obtained is just a Type I error. In that case, it is correspondingly unlikely that the effect will be replicated.

Perhaps the most disturbing conclusion from Figure 5 is that there is no way to correct the estimated aggregate replication probability for the number of opportunities without making very specific assumptions about theory strength. With an initial p value of .03 obtained after three nonsignificant pilots, for example, one might go to the appropriate panel of Figure 5 and read off the corresponding aggregate replication probability from the thickest line (i.e., "4 op."). But the researcher would have to assume either the small effect or the large effect in order to know which of these two thickest lines to use. Moreover, these lines were computed assuming that exactly 50% of the suggested effects were really present, and aggregate replication probabilities depend on this percentage as well (e.g., in Figure 4). Even when the researcher knows exactly how many opportunities there were to obtain a significant result, then, the aggregate replication probability still cannot be appropriately computed in practice, because these influential aspects of theory strength are unknown.[6]

GENERAL DISCUSSION

Two Meanings of "Replication Probability"

The results of this investigation highlight conceptual and practical complexities that researchers must face when attempting to estimate the probability of replicating an initial significant result. Conceptually, the most important problem is that "replication probability" is an ambiguous concept. It can be interpreted either as the probability of a significant result for the specific experimental effect under study ("individual" replication probability, or power) or as the probability of replicating one effect randomly selected from a large population of effects that might have been selected for study ("aggregate" replication probability). Since the numerical values of these probabilities can differ, researchers must at least decide which one they would like to estimate before attempting to determine a value.

This same conceptual distinction between two kinds of replication probability can be made for any inferential approach, not just NHST, although the precise definition of "replication" varies across approaches. In Bayesian hypothesis testing (see, e.g., Wagenmakers, 2007), for example, a researcher might consider whether an initial data set favors H_0 or H_1, and a replication could be defined as a follow-up experimental result that favors the same hypothesis as the initial experiment. Again, the individual replication probability is the probability of such a replication within the particular experimental paradigm under consideration, whereas the aggregate replication probability is the overall probability of such a replication for an experiment randomly selected from some large pool. As in the case of NHST replication probabilities, the individual replication probability would depend only on the state of the world with respect to one particular experimental paradigm, whereas the aggregate replication probability would depend on the characteristics of the whole pool of experiments from which the initial one was selected.

Another example is Killeen's (2005) p_{rep} statistic, which is an estimate of the probability that a replication will yield an effect in the same direction as the initial experiment (e.g., an experimental mean larger than the control mean). The individual p_{rep} is the probability of such a replication within the particular experimental paradigm under consideration, whereas the aggregate p_{rep} is the probability of replication across a large pool of experiments within some overall experimental context. Although Killeen (2005) did not explicitly acknowledge this distinction, he apparently intended his p_{rep} to be taken in the sense of individual replication probability, because he derived it without explicitly considering the larger research context from which the experiment was selected. Critical analyses of Killeen's (2005) p_{rep}, on the other hand, have analyzed it in both individual and aggregate terms without emphasizing the distinction between these two senses (e.g., Doros & Geier, 2005; Iverson, Lee, & Wagenmakers, 2009; Iverson, Wagenmakers, & Lee, in press). Neglecting this distinction may exacerbate the difficulty of understanding p_{rep}, especially when both senses are considered in the discussion.

Replication Probabilities Are Mostly Unknown

At a practical level, the present results illustrate several facts that must discourage researchers from attempting to estimate either of these two types of replication probability. With respect to individual replication probability, the problem is that the initial data generally provide very little information about the probability of replicating the

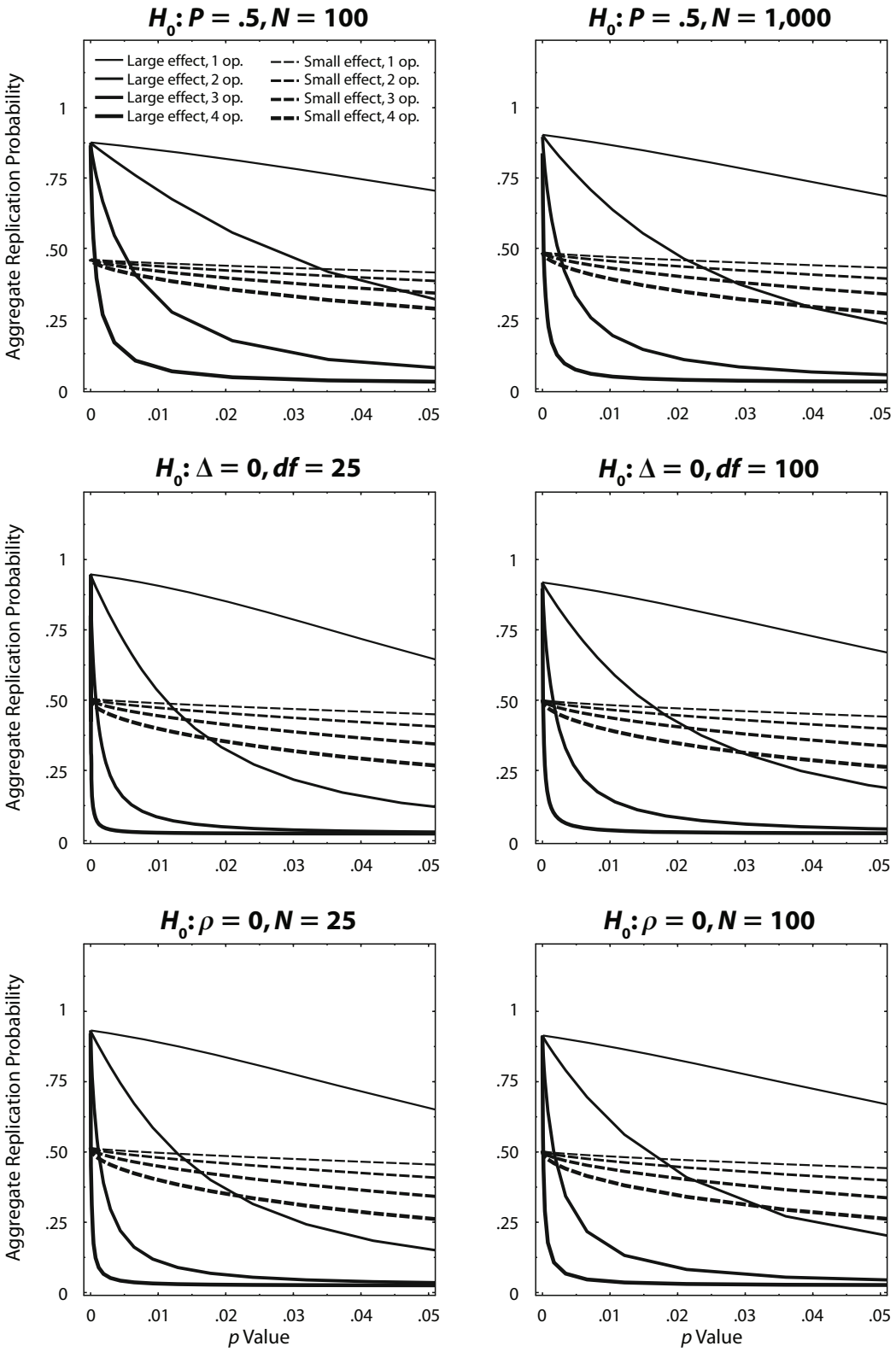

Figure 5. Aggregate replication probability as a function of the *p* value of the initial experiment, the number of opportunities for significant results (op.), the size of the true effect when it is present, and the sample size of the experiment. In all cases, real effects were assumed to be present for 50% of the null hypotheses tested. Solid lines represent theories for which the real effects are larger, whereas dashed lines represent theories for which these effects are smaller. (Top) Probability of rejecting the null hypothesis $P = .5$ using a binomial test with the indicated sample size of $N = 100$ or 1,000. (Middle) Probability of rejecting the null hypothesis that a mean or difference of means $\Delta = 0$ using a *t* test with the indicated 25 or 100 degrees of freedom (*df*) for error. (Bottom) Probability of rejecting the null hypothesis that a true correlation $\rho = 0$ for the indicated sample size of $N = 25$ or 100.

result in an identical follow-up experiment, at least when the initial result is not too highly significant (i.e., when $.005 < p < .05$). Although it is possible and convenient to estimate replication probability by assuming that the true effect matches that of the initial significant result, such an estimate ignores the statistical variability of that initial result. The true effect is almost certainly somewhat larger or smaller than was initially observed, so the true replication probability is almost certainly different from the value estimated under this assumption. Allowing for statistical variation between the observed effect and the true effect, initial results seem in most cases to be compatible with an alarmingly wide range of possible individual replication probabilities, ranging from nearly 0 in the worst case, to nearly 1 in the best (cf. Figure 3). In other words, given most sets of initial results, individual replication probability is essentially unknown, and researchers should have little confidence in any estimate of it. Perhaps this conclusion should not be surprising, given the strong dependence of power on effect size. Prior to the initial study, the researcher was not even sure whether the effect was present. Simply obtaining enough information to establish that it is present does not necessarily provide very precise information about its exact size.

With respect to aggregate replication probability, the discouraging fact is that p_{ra} depends a great deal on the strength of the theory that suggested the initial experiment. If the theory is relatively strong, in the sense that its suggested effects are mostly present and mostly large, then aggregate replication probability can be quite high. If the theory is relatively weak, suggesting effects that are mostly small or absent, this probability can be quite low. In practice, of course, researchers often work with relatively new theories, so they have little or no basis on which to judge theory strength, and this severely limits their ability to estimate aggregate replication probability. The problem is even worse when there were two or more opportunities to obtain the initial significant result. Not only does the aggregate replication probability depend on the number of opportunities, but the magnitude of this "number of opportunities" effect also depends greatly on theory strength (cf. Figure 5).

The overall practical conclusion from the present investigation of individual and aggregate replication probabilities is that researchers simply cannot expect to have a good estimate of the probability of replicating an initial significant effect in either of these two senses of "replication probability." Even though it would be very convenient to be able to estimate these replication probabilities from the initial results, too much variability and too many unknowns exist for this goal to be achievable.

Although the present article has focused on the probability of replicating a significant initial effect, very similar arguments could be made about the probability of obtaining a significant effect following a nonsignificant initial result (i.e., $p > .05$). For example, Figure 6 shows three estimates of the probability of a significant effect in a follow-up experiment after a nonsignificant initial result, as a function of the p value in the initial experiment. As in Figure 3, the three probability estimates were computed both from the observed effect and from the upper and lower bounds of a 95% confidence interval for the true effect, so they are analogous to individual replication probabilities. What is perhaps surprising is that the probability of a significant effect can be quite high in the follow-up—around .5—even if there was virtually no effect at all in the initial experiment (i.e., p level near 1.0). Again, the wide range from the lowest to the highest probabilities simply reflects the statistical uncertainty regarding the size of the true effect.

The uncertainties highlighted in this article also have implications concerning the use of pilot studies to obtain power estimates, which are analogous to individual replication probabilities. Researchers embarking on a new line of experimentation may want to conduct small pilot studies to get initial estimates of effect size on which to base power calculations for a main study that is to follow. The temptation to do this is certainly increased by exhortations to compute post hoc power (e.g., Greenwald et al., 1996; Onwuegbuzie & Leech, 2004) and by the availability of convenient software for performing the relevant calculations (e.g., SPSS, 2006). The present results suggest, however, that such pilot studies will rarely be useful for this purpose, because they will not provide narrow enough constraints on the effect size to constrain power very much. Although one can of course use a pilot study's observed effect size to estimate power, the resulting estimated power value could be quite misleading because of statistical error in the effect size estimate (cf. Figures 3 and 6). There are many reasons for carrying out pilot studies, but obtaining an estimated effect size for power computations would not appear to be one of them. Researchers who do want to estimate power from pilot studies should at least compute a confidence interval for the true effect size and then compute best-case and worst-case power estimates corresponding to the boundaries of this interval.

Meta-Analysis and Aggregate Replication Probability

Given the importance of theory strength for aggregate replication probability, one might try to assess this strength empirically to improve estimates of p_{ra}. With sufficient literature reviews and meta-analyses (e.g., Cohen, 1962; Lipsey & Wilson, 1993; Richard, Bond, & Stokes-Zoota, 2003), it might eventually be possible to formulate reasonable assumptions about a theory's strength. For example, one might estimate empirically what proportion of significant initial results in an area turn out to be replicated. As attractive as this approach might seem, the problem of selection artifacts creates significant obstacles for it. For example, the bias against publication of nonsignificant results tends to mean that published effects are, on average, larger than true effects (see, e.g., Rosenthal, 1979). Such bias works to make the theories predicting these effects appear stronger than they actually are. As a further example, consider the biases that would influence researchers selecting an effect for study in a meta-analysis. No really large effect would be selected, because such effects are so easy to establish that meta-analysis is unnecessary. Furthermore, it is unlikely that many truly nonexistent effects would be selected either. Such effects are by definition only observed

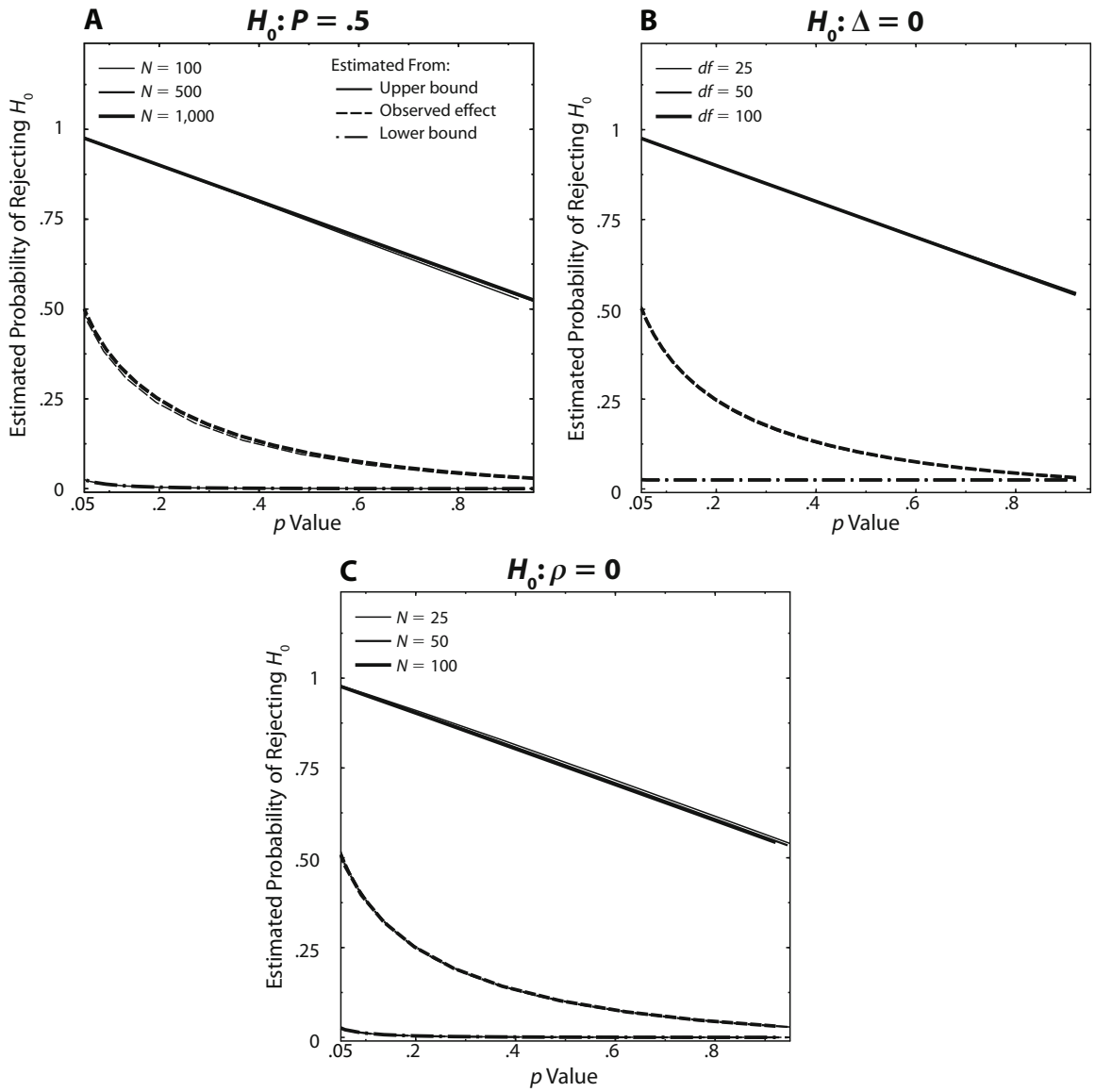

Figure 6. Three estimates of the probability of rejecting the null hypothesis in an identical replication experiment as a function of the *p* value of a nonsignificant initial experiment (i.e., $p > .05$). The format of this figure is identical to that of Figure 3, except that the range of the horizontal axis corresponds to nonsignificant initial results.

significantly at the Type I error rate. The many negative results emerging from initial investigations of these effects would discourage researchers from undertaking enough additional studies to support a meta-analysis. Thus, it seems clear that meta-analyses will tend to examine effects that are small yet real rather than examining a representative cross-section of all effects that have been studied. Because of such biases in the selection of effects for meta-analysis, no survey of existing meta-analyses can be used to make inferences about the likely size of all experimental effects. These influences of sampling bias on the topics chosen for meta-analysis are easy to overlook. Hunter (1997), for example, noted that almost all published meta-analyses—across a wide range of research domains—have concluded that the effect under study was real (Lipsey & Wilson, 1993). From that fact, he concluded that "empirical studies have now shown that the null hypothesis is rarely true" (Hunter, 1997, p. 5). This conclusion is not valid, however, because these meta-analyses checked a biased set of null hypotheses (i.e., those receiving meta-analyses). The null hypotheses may have been true for quite a few effects that received too little empirical investigation to warrant meta-analyses.

Implications for NHST

What are the implications of uncertain replication probability for the wider debate concerning the utility of NHST as a statistical tool (e.g., Estes, 1997; Fraley & Marks, 2007; Krueger, 2001; Morgan, 2003; Nickerson, 2000)? NHST has previously been criticized because *p* values are

"very unreliable" (Cumming, 2008, p. 286); that is, they tend to vary widely across identical replications of a given experiment. In addition, it now appears that the *p* value provides relatively little information about the probability of replication. It is certainly tempting to conclude that these uncertainties indicate deep flaws within NHST itself (e.g., Cumming, 2008; Sohn, 1998; Thompson, 1996).

As compelling as these criticisms may appear, it is actually quite unfair to regard them as indicative of a problem unique to NHST. The unreliability of *p* values and the uncertainty of replication probabilities arise purely out of sampling error, not out of flaws in NHST. After all, *p* values and estimated replication probabilities are merely statistics computed from the overall set of experimental results. These quantities vary across replications precisely because the overall results vary, and this same sampling variability creates comparable uncertainty regarding the measures associated with every other inferential technique.

Figure 7 illustrates this argument concretely using the example of a binomial experiment with 100 trials. Panel A illustrates the sampling distribution of the most basic summary measure of the experimental outcome—the observed number of successes, *i*. The probability of each outcome *i* depends on the true probability of success, *P*, as is illustrated by the different sampling distributions for $P = .51$ and $P = .59$. Each distribution is plotted as a series of points to emphasize the fact that there is a discrete set of possible experimental outcomes, $i = 0, 1, \ldots, 100$. It is clear from these probability distributions that the observed number of successes can vary widely across identical replications (i.e., with a fixed value of *P*).

Panel B shows the corresponding sampling distributions for another summary measure of the experimental results—namely, the *p* value obtained when testing the null hypothesis H_0: $P = .5$. Each experiment's *p* value depends only on its observed number of successes, so there is a one-to-one mapping from the points in panel A to the points in panel B, as is illustrated by the arrow showing the *p* value computed for $i = 44$.[7]

Inspection of panel B shows that *p* values do vary widely across replications of an identical experiment. With $P = .51$, for example (open squares), a given replication might yield a *p* value approaching either of the two extreme possibilities (i.e., 0 or 1). This variability clearly shows the unreliability of *p* values that was emphasized by Cumming (2008). Critically, however, the one-to-one mapping from the points in panel A to the points in panel B makes it obvious that the variability of *p* values is a direct reflection of the underlying variation in the observed numbers of successes. In short, the problem of unreliable *p* values is simply another manifestation of the problem of sampling variability. Analogous unreliability necessarily plagues every measure computed from the relevant sample statistics, as is illustrated in panels C–F, so this unreliability is in no way peculiar to NHST.

Panels C and D illustrate analogous variability in the estimated individual and aggregate replication probabilities, p_{ri} and p_{ra}, respectively. As was discussed in the Estimation of Individual Replication Probability section, individual replication probability is estimated from each experimental outcome by assuming that the true population proportion *P* exactly matches the observed value (i.e., *i*/100), so each observed value in panel A corresponds to one estimated p_{ri}. As noted in the earlier section, these p_{ri} values are the same as post hoc power values (e.g., Onwuegbuzie & Leech, 2004) and observed power values (SPSS, 2006), and they can be computed whether the result of the initial experiment is significant or not. Panel D shows aggregate replication probability, which is estimated from the observed results using Equation 8, in this case for an arbitrary example with equally likely prior probabilities of the alternative hypotheses H_0: $P = .5$ and H_1: $P = .6$.

The important point illustrated by panels C and D is that both types of estimated replication probability vary substantially from one replication to the next, which further illustrates the fact that neither true replication probability can be estimated very precisely from the results of a single experiment. Again, however, the uncertainty about these replication probabilities can be traced directly back to the sampling variability of *i*, because there is a one-to-one mapping from the points in panel A to those in panels C and D. Thus, the uncertainty associated with p_{ri} and p_{ra} also reflects the variability inherent in experimental results rather than a flaw specific to NHST.

Finally, panels E and F illustrate equivalent variation in two other kinds of summary measures that have been suggested as improvements over NHST's *p* values. Panel E shows variation in a binomial analogue of Killeen's (2005) p_{rep}. As mentioned earlier, Killeen (2005) considered differences in experimental versus control group means, and he defined the probability of replication, p_{rep}, in terms of the probability of replicating an effect of a particular direction of effect (i.e., whether the control or experimental mean was larger). For a binomial experiment designed to assess H_0: $P = .5$, an analogous p_{rep} can be defined as the probability of more than 50% successes, and this can be estimated by assuming that the true probability matches the observed proportion of successes (i.e., *i*/100), as was done for panel C. As with all of the other summary measures, p_{rep} estimates vary substantially across replications because of variation in the observed number of successes from which they are computed.[8]

Similarly, panel F shows variation in the Bayes factor (see, e.g., Kass & Raftery, 1995), another measure that has been suggested as an improvement over *p* values (e.g., Wagenmakers, 2007; see also Glover & Dixon, 2004, for a similar proposal). In brief, the Bayes factor is a measure of the relative likelihood of the data set under each of two competing hypotheses, chosen arbitrarily for this example to be H_0: $P = .5$ versus the particular H_1: $P = .6$. A log Bayes factor of ± 1, for example, means that the data are $10^1 = 10$ times more likely under one hypothesis than the other, and Bayesians would consider this positive evidence for the favored hypothesis (e.g., Raftery, 1995). The Bayes factor is also computed directly from the number of successes, so it too varies across replications in a manner deriving entirely from the variation of *i* (panel A). With a true *P* of .51, for example, one might observe 49 successes

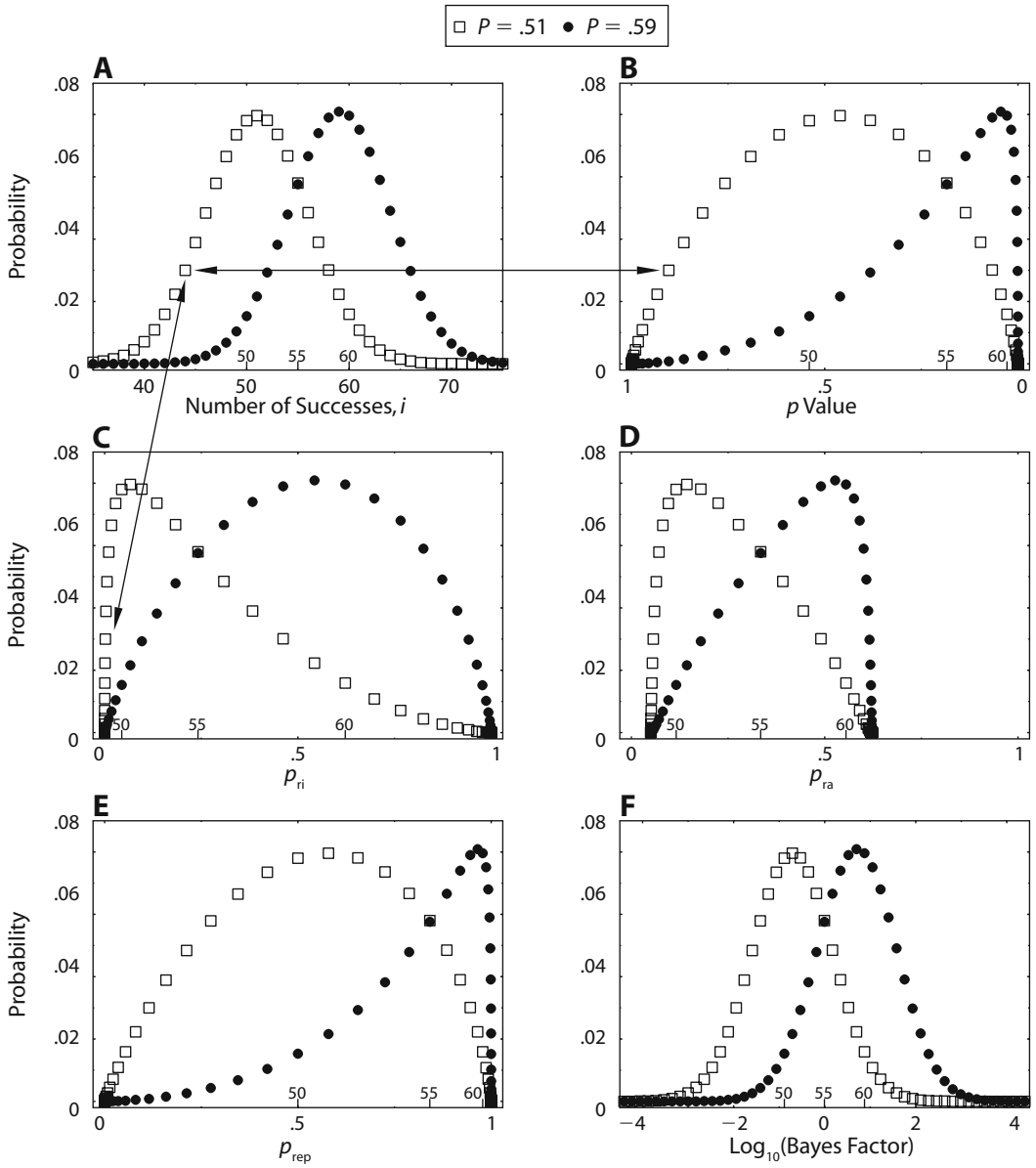

Figure 7. The sampling distributions of various outcome measures that could be used to summarize the results of a binomial experiment with 100 trials. Each distribution is shown as a series of discrete points to emphasize the fact that such an experiment has discrete outcomes corresponding to different whole numbers of successes (i.e., 0–100). The probability of each outcome is shown on the vertical axis in all panels, separately for examples in which the true probability of success is $P = .51$ or .59. (A) Sampling distributions of the number of successes, i. The sampling distributions in all of the other panels reflect this same discrete set of outcomes, with a one-to-one mapping between the points for every pair of panels. For example, the arrows linking panel A to panels B and C illustrate this mapping by linking corresponding points associated with the outcome $i = 44$. To facilitate the visualization of the one-to-one mapping across all panels, the outcomes corresponding to $i = 50$, 55, and 60 successes are indicated along the horizontal axis of each panel. (B) Sampling distributions of p values for tests of the null hypothesis $P = .5$, one-tailed, against the alternative that $P > .5$. Note that the horizontal axis has been reversed so that the most significant results are shown at the far right, corresponding to the largest i values. (C) Sampling distributions of estimated individual replication probability, p_{ri}. Specifically, p_{ri} is the estimated probability of rejecting H_0: $P = .5$ ($\alpha = .05$, one-tailed), assuming that the true P value corresponds exactly to the observed number of successes, $P = i/100$. (D) Sampling distributions of estimated aggregate replication probability, p_{ra}, computed using Equation 8 under the assumption of two equally likely alternative hypotheses H_0: $P = .5$ versus H_1: $P = .6$. (E) Sampling distribution of an estimated probability of replication analogous to that proposed by Killeen (2005), p_{rep}. p_{rep} is the probability of observing more than 50 successes, plus half of the probability of observing exactly 50 successes, under the assumption that the true P is the observed proportion of successes $i/100$. (F) Sampling distributions of the Bayes factor. The Bayes factor is the ratio of the likelihoods of the observed i value under two competing hypotheses, H_1: $P = .6$ versus H_0: $P = .5$, plotted on a log scale to improve its appearance.

and conclude there was positive evidence in favor of H_0 (i.e., log Bayes factor = −1.06), or one might observe 61 successes and conclude there was positive evidence in favor of H_1 (log Bayes factor = 1.05). Thus, Bayes factors are also rather variable, just like the NHST-based measures, because of the sampling error associated with the underlying number of successes.

In summary, the main lesson illustrated by Figure 7 is that all measures derived from overall experimental outcomes are subject to the natural variability inherent in those outcomes. It is unfair to single out NHST for the criticism that its *p* values, replication probabilities, and other measures are "unreliable"; these measures are no more unreliable than the experimental outcomes from which they are computed. Exactly corresponding variability is present in other inferential approaches (e.g., Bayes factors), so the uncertainties associated with replication probabilities have no direct implications for the controversy regarding which inferential approach is best. In fact, the one-to-one correspondence of each measure to the number of successes, *i*, implies that the measures in the different panels are all monotonically related to one another within a given experimental design. Thus, even though these measures might have numerically different variances, the measures all provide exactly the same information about the experimental outcome, in the technical "information transmission" sense of Shannon and Weaver (1949). Note, for example, that the curves for the true success proportions of $P = .51$ and $P = .59$ intersect at the same experimental outcome in all panels (i.e., the outcome corresponding to $i = 55$), so all measures yield equivalent partitions of the outcomes into those favoring $P = .51$ versus those favoring $P = .59$. In this sense, the different ways of summarizing a given set of data are all informationally equivalent.

The informational equivalence of the different measures in Figure 7 has not always been fully acknowledged. In discussing Killeen's (2005) p_{rep}, for example, Wagenmakers (2007) noted correctly that "the p_{rep} statistic can be obtained from the NHST *p* value by a simple transformation" (p. 780). He seemed to conclude from this equivalence, however, that "p_{rep} inherits all of the *p* value problems" (p. 780) and therefore must be flawed. Also considering Killeen's (2005) work, Doros and Geier (2005) similarly commented that "any measure that is no more than a simple transformation of the classical *p* value (see Killeen's [2005] appendix) will inherit the shortcomings of that *p* value" (p. 1006). By such reasoning, however, every other measure derived from *i* would also inherit these problems, including not only p_{rep} but also the Bayesian measures (e.g., panel F) often put forth as an improvement.

Indeed, this informational equivalence may be quite surprising, because we intuitively expect more sophisticated methods of statistical analyses to be more successful in overcoming variability. Unfortunately, even the best statistical methods have limited resolution because of the variability of the basic data being analyzed. Ultimately, variability must be overcome by increasing sample size and reducing measurement error, not by improving statistical techniques. Different inferential methods may be especially well suited to answering different questions, though. For example, NHST is oriented toward assessing the plausibility of a particular (null) hypothesis in isolation, whereas Bayesian methods emphasize comparisons among two or more alternative hypotheses. Most standard methods are based on the same underlying sample summary values, however (i.e., the so-called "sufficient" statistics that capture all of the relevant information available in the sample), so they are equally sensitive to sampling variability at the point of drawing inferences from new data.

Precise *p* Values Versus Unknown Replication Probabilities—A Double Standard?

Given the sampling variability present in all inferential measures used to summarize experimental outcomes, it may seem strange that inferential techniques are based on very precise assessments of the particular observed experimental results. When using NHST, for example, researchers reject the null hypothesis if the *p* value is .049 but not if it is .051, despite the obvious possibility that a replication would yield a result on the other side of the .05 borderline. Similarly, a researcher using Bayesian methods might conclude that the data increase the likelihood of H_1 relative to H_0 by a factor of precisely (say) 10.43, despite the possibility that a replication could easily counteract that evidence entirely, favoring H_0 over H_1 by the same margin. Even confidence intervals, which explicitly acknowledge the uncertainty involved in generalizing from samples to populations by providing a range of possible values for a population parameter, are computed precisely. For example, the confidence interval for a mean might cover the precise range 287.3–423.7. What justifies the computation of such precise values from such variable results? Are we, as one reviewer suggested, applying a "double standard" when we claim to know the *p* value precisely but to have little idea of the replication probabilities p_{ri} and p_{ra}?

Not at all. Precise calculations from experimental results, including *p* values, summarize the information obtained from the sample, which is indeed known with great precision. In a given experiment with 100 binomial trials, for example, the obtained number of successes in that sample is observed exactly. From that number, the exact *p* value associated with that sample can also be computed, as can exact values for p_{rep}, the Bayes factor, and any other summary of the results in that sample. The *p* value, for example, does indicate precisely how unusual the particular observed data are, given the predictions of the null hypothesis. Similarly, the Bayes factor indicates precisely how likely those particular data are under each of two alternative hypotheses. Thus, the *p* value, the Bayes factor, and other summary measures associated with the current sample can be determined exactly from the results already collected, just as the observed number of successes is known exactly.

Unlike the characteristics of the sample, the characteristics of the population are not known exactly once the

sample has been observed, and this is why p_{ri} and p_{ra} are not known exactly even though p is. Replication probabilities are used to predict what will happen with future samples, so they require knowledge about the whole population—not just about the current sample. Replication probabilities can be estimated from a sample, just like population proportions or means, but the error associated with these estimates may be quite large (see, e.g., Figure 3). Thus, precise calculations with sample data lead to imprecise conclusions about populations. This does not reflect a double standard, but instead simply reflects the researcher's different levels of knowledge about the sample versus population.

AUTHOR NOTE

I thank Joachim Krueger, Wolf Schwarz, Rolf Ulrich, Esther Vierck, Eric-Jan Wagenmakers, and an anonymous reviewer for constructive comments on earlier versions of the article. Correspondence concerning this article should be addressed to J. Miller, Department of Psychology, University of Otago, Dunedin, New Zealand (e-mail: miller@psy.otago.ac.nz).

REFERENCES

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, **8**, 12-15.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, **32**, 1369-1378.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking & Learning*, **2**, 75-97.

Boynton, D. M. (2003). Superstitious responding and frequency matching in the positive bias and gambler's fallacy effects. *Organizational Behavior & Human Decision Processes*, **91**, 119-127.

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

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155-159.

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

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, **16**, 1002-1004.

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

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

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, **60**, 170-180.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, **11**, 217-227.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, **3**, 299-311.

Doros, G., & Geier, A. B. (2005). Probability of replication revisited: Comment on "An alternative to null-hypothesis significance tests." *Psychological Science*, **16**, 1005-1006.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science*, **8**, 18-20.

Falk, R. (1998). Replication—A step in the right direction: Commentary on Sohn. *Theory & Psychology*, **8**, 313-321.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, **39**, 175-191.

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance-testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149-169). New York: Guilford.

Froman, T., & Shneyderman, A. (2004). Replicability reconsidered: An excessive range of possibilities. *Understanding Statistics*, **3**, 365-373.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M., & Greenberg, D. A. (2007). Non-replication of association studies: "Pseudo-failures" to replicate? *Genetics in Medicine*, **9**, 325-331.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology*, **33**, 175-183.

Guttman, L. (1977). What is not what in statistics. *Statistician*, **26**, 81-107.

Haller, H., & Kraus, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, **7**, 1-20.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science*, **8**, 8-11.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, **55**, 19-24.

Hogg, R. V., & Craig, A. T. (1970). *Introduction to mathematical statistics* (3rd ed.). New York: Macmillan.

Huberty, C. J., & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 1-22). Stamford, CT: JAI Press.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, **8**, 3-7.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, **2**, 696-701.

Iverson, G. J., Lee, M. D., & Wagenmakers, E.-J. (2009). p_{rep} misestimates the probability of replication. *Psychonomic Bulletin & Review*, **16**, 424-429.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (in press). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345-353.

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, **13**, 549-562.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, **56**, 16-26.

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

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, **5**, 161-171.

Lykken, D. T. (1991). What's wrong with psychology, anyway? In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl. Matters of public interest* (Vol. 1, pp. 2-39). Minneapolis: University of Minnesota Press.

Macdonald, R. R. (2003). On determining replication probabilities: Comments on Posavac (2002). *Understanding Statistics*, **2**, 69-70.

MACDONALD, R. R. (2005). Commentary: Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, **16**, 1007-1008.
MORGAN, P. L. (2003). Null hypothesis significance testing: Philosophical and practical considerations of a statistical controversy. *Exceptionality*, **11**, 209-221.
MURTY, V. N., & BISSINGER, B. H. (1982). The law of succession and Bayes' rule. *Two-Year College Mathematics Journal*, **13**, 44-51.
NEWCOMBE, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, **17**, 857-872.
NICKERSON, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, **5**, 241-301.
OAKES, M. L. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
ONWUEGBUZIE, A. J., & LEECH, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, **3**, 201-230.
POSAVAC, E. J. (2002). Using *p* values to estimate the probability of a statistically significant replication. *Understanding Statistics*, **1**, 101-112.
PSYCHOLOGICAL SCIENCE EDITORIAL BOARD (2005). Information for contributors. *Psychological Science*, **16**(12).
RAFTERY, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111-196). Cambridge, MA: Blackwell.
RICHARD, F. D., BOND, C. F., JR., & STOKES-ZOOTA, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, **7**, 331-363.
ROBINSON, D. H., & LEVIN, J. R. (1997). Research news and comment: Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, **26**, 21-26.
ROSENTHAL, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, **86**, 638-641.
ROSENTHAL, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Erlbaum.
SHAFFER, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561-584.
SHANNON, C., & WEAVER, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
SOHN, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, **8**, 291-311.
SPSS INC. (2006). SPSS 14.0 for Windows [Computer software]. Chicago: Author.
SUNDALI, J., & CROSON, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment & Decision Making*, **1**, 1-12.
THOMPSON, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, **25**, 26-30.
THOMPSON, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, **31**, 25-32.
TUKEY, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, **24**, 83-91.
TVERSKY, A., & KAHNEMAN, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.
WAGENMAKERS, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, **14**, 779-804.
WAINER, H., & ROBINSON, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, **32**, 22-30.
WILSON, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

NOTES

1. The exact definition of the true effect size depends on the null hypothesis being tested, and this article considers three possibilities as examples. One uses the binomial test to assess the null hypothesis that $P = .5$, where P is the true probability of success. For this test, the true effect size is simply the difference $P - .5$. A second uses a t test to check the null hypothesis that $\Delta = 0$, where Δ is either a single mean (i.e., one-sample t test), a difference between the means of two conditions with paired data (i.e., paired t test), or a difference between the means of independent groups (i.e., two-sample t test). For this test, the true effect size is Δ/σ, where σ is a measure of the population standard deviation of the scores or difference scores being compared. The third test checks the null hypothesis that $\rho = 0$, where ρ is the true population-wide correlation between two variables. For this test, the effect size is simply the true value of ρ.

2. Other uses of the term "replication" have also been proposed. For example, Rosenthal (1993) suggested that much more weight be given to effect sizes than to significance levels. In contrast, Killeen (2005; see also Cumming, 2005) proposed that it might be more useful to consider an effect to have been replicated if the follow-up study obtained results in the same direction as the initial study, without regard to statistical significance. Because these proposals have not yet gained wide acceptance and NHST is still in common use, however, in this article "replication," for better or worse, will generally be used in the traditional sense requiring statistical significance, with extensions to some other meanings considered briefly.

3. For simplicity, in this article I consider only the probability of replicating an initial significant result in an identical follow-up experiment (i.e., same population and sample size, measurement error, etc.), although in principle the same considerations arise in replications with different sample sizes, error variances, and so on.

4. One might try to extend this argument about relative p values to obtain an estimate of individual replication probability. Specifically, considering a single significant result in isolation, one might estimate that the individual replication probability is always at least .5, since the replication is just as likely to be "more significant" than the initial experiment as it is to be "less significant." As will be considered later, though, it is not always appropriate to consider a single experiment in isolation. Moreover, even when it is appropriate, one must keep in mind that this estimated individual replication probability of at least .5 is only an estimated value based on observed data—not a true value.

5. This law emerges from a Bayesian analysis in which the probability of success is assumed to have a uniform prior distribution across the 0–1 interval (for a derivation, see, e.g., Murty & Bissinger, 1982).

6. Although it might seem that the researcher would know the number of opportunities exactly, this need not always be true in practice. When researchers carry out a series of pilot studies, they often make various procedural changes from one to the next. If these changes affect power, the "number of opportunities" does not increase incrementally with each successive pilot study. Since the researcher cannot know the exact power level of each study, it is impossible to be sure exactly how the number of opportunities depends on the number of such studies.

7. Corresponding points have identical values on the ordinates in the two panels but have different values on the abscissas. Note that the distribution of p values is also discrete, like that of the number of successes, even though the p values are not whole numbers.

8. The visual similarity of panels B and E is rather striking. Although it is beyond the scope of this article to investigate the precise relationship between traditional p values and Killeen's (2005) p_{rep}, I did find that p_{rep} and $1-p$ were approximately equal under many scenarios. This same relationship is also evident in Equation 3 of Doros and Geier (2005) and in Figure 1 of Iverson, Wagenmakers, and Lee (in press).

APPENDIX A
Computation of Aggregate Replication Probability

This appendix presents the assumptions and formulas used for the computation of the aggregate replication probabilities shown in Figures 4 and 5. Two numerical examples illustrating the computations are shown in Table A1.

Table A1
Numerical Examples for Computation of Aggregate Replication Probability (p_{ra})

Measure	Hypothesis	
	H_0: $P = .5$	H_1: $P = .6$
Pr(Reject H_0 & conclude $P > .5$)	.0176	.4621
Prior Pr(H)	.75	.25
Example 1: $O = 62$ Successes		
Pr($O \mid H$)	.0045	.0754
Pr($H \mid O$)	.1511	.8489
p_{ra}	$.0176 \times .1511 + .4621 \times .8489 = .3949$	
Example 2: Observe Three Pilots Retaining H_0, and Then $O = 62$ Successes		
Pr($3 \mid H$)	.8981	.1556
Pr($O \cap 3 \mid H$)	.0040	.0117
Pr($H \mid O \cap 3$)	.5067	.4933
p_{ra}	$.0176 \times .5067 + .4621 \times .4933 = .2369$	

Note—The two examples illustrate calculations for binomial experiments with 100 trials. According to the null hypothesis (H_0), the true probability of success is $P = .5$; according to the alternative hypothesis (H_1), this probability is $P = .6$. The prior probabilities of the null and alternative hypotheses are .75 and .25, respectively. The null hypothesis is rejected and it is concluded that $P > .5$ if 61 or more successes are observed. In Example 1, the observed data consist of a single experiment resulting in $O = 62$ successes. Pr($O \mid H$) is the probability of this observed result under each hypothesis, computed with the binomial formula. Pr($H \mid O$) is the corresponding posterior probability of each hypothesis, computed using Bayes's theorem. The aggregate replication probability p_{ra} is computed from these values using Equation 8. Example 2 is analogous to Example 1, except that the observed data consist of three pilot experiments in which the null hypothesis was retained followed by a fourth experiment with $O = 62$ successes.

For simplicity, I only studied the effects of research context using theories for which predicted effects are present, with probability γ, or absent, with probability $1-\gamma$. The true size of each effect was adjusted to match the observed size that would produce a given p value of .05 or .001. Let Pr($O \mid H_0$) and Pr($O \mid H_1$) denote the probabilities (or probability densities) of any given observed initial results, O, under the null and alternative hypotheses, respectively. Then, according to Bayes's theorem (e.g., Hogg & Craig, 1970),

$$\Pr(H_1 \mid O) = \frac{\Pr(H_1) \cdot \Pr(O \mid H_1)}{\Pr(H_1) \cdot \Pr(O \mid H_1) + \Pr(H_0) \cdot \Pr(O \mid H_0)} \tag{A1}$$

Aggregate Replication Probability

Given a randomly selected significant initial result, the conditional probability that it will be replicated in the same direction in one follow-up experiment is

$$p_{ra} = \frac{\alpha}{2} \Pr(H_0 \mid O) + (1 - \beta) \cdot \Pr(H_1 \mid O), \tag{A2}$$

where $\alpha/2$ is the probability of a significant result in the same direction under the null hypothesis[A1] and $1-\beta$ is the power of the experiment when the effect is present (i.e., the probability that it will yield a significant result).

Effect of Number of Opportunities

To study the influence of the number of opportunities, Equation A1 must be elaborated to include k nonsignificant experimental results prior to the initial significant observed result, O. This can be done by conceiving of the data as the full set of observations, including all k nonsignificant results as well as the observed significant one, O. Under the null hypothesis, the probability of the set of k nonsignificant results is Pr($O \mid H_0$) $\cdot$ $[1-\alpha]^k$; under the alternative hypothesis, it is Pr($O \mid H_1$) $\cdot$ $[\beta]^k$. These values may be used to compute the probability of H_1 given the full set of results via Bayes's theorem:

APPENDIX A (Continued)

$$\Pr\left(H_1 \mid O \cap k\right) = \frac{\Pr\left(H_1\right) \cdot \Pr\left(O \mid H_1\right) \cdot \left[\beta\right]^k}{\Pr\left(H_1\right) \cdot \Pr\left(O \mid H_1\right) \cdot \left[\beta\right]^k + \Pr\left(H_0\right) \cdot \Pr\left(O \mid H_0\right) \cdot \left[1 - a\right]^k}. \quad \text{(A3)}$$

This value may be used in place of $\Pr(H_1 \mid O)$ in Equations A2 and A4 when computing the aggregate replication probability after a series of k nonsignificant results followed by a significant one.

Theory Strength and the Probability of Multiple Replications

The results shown in Figure 4 concern the aggregate probability of obtaining a single significant replication, but analogous probabilities can also be computed for the probability of $j = 2, 3, \ldots$ replications. These aggregate multiple replication probabilities would be especially relevant to a researcher planning a series of studies—for example, when preparing a grant proposal.

How should the probability of j successful replications be computed? Initially, it might seem that if the probability of a single replication is p_{ra}, the estimated probability of j replications should be p_{ra}^j. The situation is somewhat more complicated, however, when theory strength is taken into account, because the different replications are not all independent. Instead, they all depend in the same way on the theory that suggested the initial experiment, as the following analysis illustrates. Consequently, the probability that a randomly selected significant initial effect will be replicated in j consecutive follow-up experiments is

$$p_{\text{ra}} = (\alpha/2)^j \Pr(H_0 \mid O) + (1-\beta)^j \cdot \Pr(H_1 \mid O). \quad \text{(A4)}$$

As extreme examples, consider researchers A and B working with two different theories. Researcher A's theory predicts only real effects (i.e., $\gamma = 1$), but the effects are only moderate in size, so any given effect is detected with a power of $1-\beta = .5$ in a typical experiment. For this researcher, there is a 50% chance that any given replication attempt will succeed, so the probability that j successive replication attempts will all succeed is $p_{\text{ra}} = .5^j$.

For researcher B, in contrast, only half of the theoretically predicted effects are real (i.e., $\gamma = .5$), but these effects are so large that experimental power is virtually 100% (i.e., $1-\beta = 1$). This means that the probability of a single successful replication attempt is approximately .5 for researcher B, just as it was for researcher A. For researcher B, though, the probability of j successful replications does not decline much as j increases, unlike the situation with researcher A, because any real effect is virtually always replicable.

Figure A1 illustrates that the aggregate probability of j successful replications can vary with j for the same scenarios illustrated in Figure 4. The computations were analogous to those presented in Figure 4 but were based on Equation A4, assuming in all cases that the suggested effect was real with probability = .5. With a small effect, the aggregate probability of j successful replications is quite close to $.5^j$, because experimental power is approximately .5 for this real effect. With a large effect, though, the aggregate probability of j successful replications decreases rather slowly as j increases; if the large effect is real, it is replicable with quite high probability.

Figure A1 is important because it illustrates further the complexity of the concept of aggregate replication probability. Even with a very specific hypothetical scenario under which the probability of one replication can be estimated, the estimated probability of j "independent" replications is not simply p_{ra}^j.

NOTE

A1. Typically, this is .025 for a two-tailed test with an α level of .05. For tests of proportions, however, the desired α and $\alpha/2$ levels cannot be attained exactly, because of the discreteness of the observed number of successes. In those cases, the computations used whatever slightly smaller values of α and $\alpha/2$ were possible.

APPENDIX A (Continued)

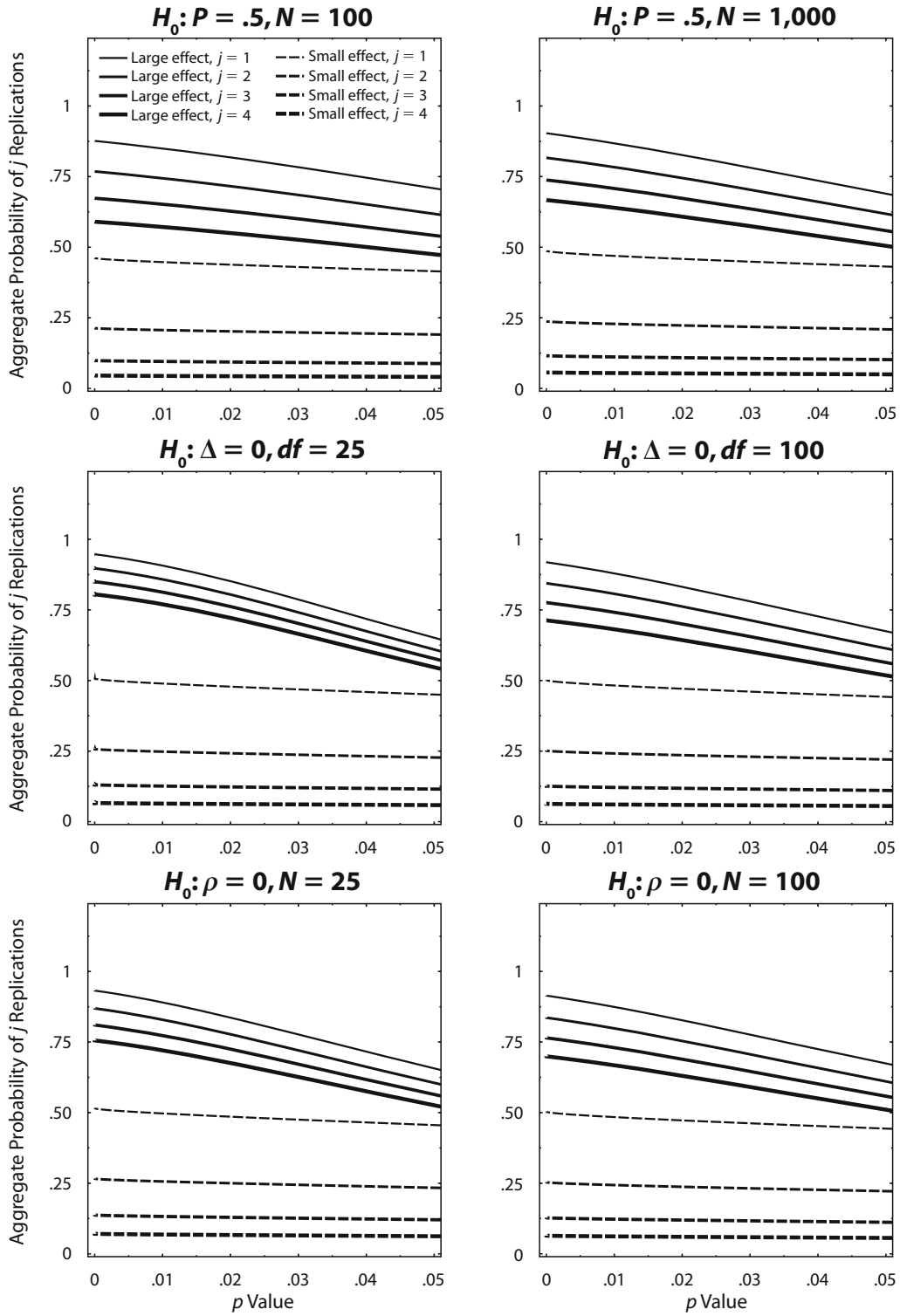

Figure A1. Aggregate replication probability as a function of the number of successful replications sought (*j*), the *p* value of the initial experiment, and the strength of the background theory on which the initial experiment was based. The line thickness represents the number of successful replications that are sought. Solid lines represent theories for which the real effects are larger, whereas dashed lines represent theories for which the effects are smaller. In all cases, real effects were assumed to be present for 50% of the null hypotheses tested. (Top) Probability of rejecting the null hypothesis $P = .5$ using a binomial test with the indicated sample size of $N = 100$ or 1,000. (Middle) Probability of rejecting the null hypothesis that a mean or difference of means $\Delta = 0$ using a *t* test with the indicated 25 or 100 degrees of freedom (*df*) for error. (Bottom) Probability of rejecting the null hypothesis that a true correlation $\rho = 0$ for the indicated sample size of $N = 25$ or 100.

APPENDIX B
Computation of Confidence Intervals

For ease of computation, confidence intervals for proportions, means, and correlations have traditionally been computed as an observed value plus or minus a half-width, with normal approximations used to establish the half-width for proportions and correlations. This traditional approach, which is most closely associated with the interpretation of a confidence interval as having a 95% (say) probability of including the true value (cf. Cumming & Finch, 2001, Method 1), was used for the computations reported here. Virtually identical results were obtained, however, with an alternative approach in which the confidence interval limits were determined by adjusting the parameter of the binomial, noncentral *t*, or noncentral correlation distribution to determine a range of parameter values that cannot be rejected by a standard significance test (cf. Cumming & Finch, 2001, Method 2; Thompson, 2002).

Proportions

For a binomial test with N trials and i observed successes, a traditional 95% confidence interval for the true proportion of successes P can be obtained using the normal approximation to the binomial (see, e.g., Newcombe, 1998):

$$\hat{p} - \text{HW} \leq P \leq \hat{p} + \text{HW}, \tag{B1}$$

where $\hat{p} = i/N$ and $\text{HW} = 1.96\sqrt{\hat{p}(1-\hat{p})/N}$. Other techniques for computing confidence intervals for proportions have been proposed and may have better coverage properties (e.g., Agresti, 2002), but the simpler and more traditional binomial approximation was used here because it will be familiar to most readers.

t Tests

Assume for simplicity and without loss of generality that the standard error of the sample mean equals 1. In that case, for a t test with df degrees of freedom, a 95% confidence interval for the true mean or difference in means, Δ, is

$$t_o - \text{HW} \leq \Delta \leq t_o + \text{HW}, \tag{B2}$$

where t_o is the obtained t value and HW is the (97.5)th percentile point of Student's t distribution with df degrees of freedom.

Correlations

With an observed correlation r based on N cases, an approximate 95% confidence interval for the true correlation ρ can be obtained using Fisher's r-to-z transformation:

$$z_r = .5[\ln(1 + r) - \ln(1 - r)] \tag{B3}$$

(see, e.g., Hays, 1988). The sampling distribution of z_r is approximately normal with $\sigma_{z_r} = 1/\sqrt{N-3}$, so confidence interval bounds for z_ρ are

$$z_r - \text{HW} \leq z_\rho \leq z_r + \text{HW}, \tag{B4}$$

where $\text{HW} = 1.96/\sqrt{N-3}$. The upper and lower confidence limits for ρ can then be found from the upper and lower bounds for z_r by inverting the transformation of Equation B3:

$$r_z = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}. \tag{B5}$$

(Manuscript received October 16, 2008;
revision accepted for publication February 20, 2009.)