Psychological Methods
2010, Vol. 15, No. 3, 300–307

1082-989X/10/$12.00 DOI: 10.1037/a0018533

How Often Is p_{rep} Close to the True Replication Probability?

David Trafimow, Justin A. MacDonald, Stephen Rice, and Dennis L. Clason
New Mexico State University

Largely due to dissatisfaction with the standard null hypothesis significance testing procedure, researchers have begun to consider alternatives. For example, Killeen (2005a) has argued that researchers should calculate p_{rep} that is purported to indicate the probability that, if the experiment in question were replicated, the obtained finding would be in the same direction as the original finding. However, Killeen also seems to indicate that rather than being the probability of replication, p_{rep} is actually the probability of obtaining a finding whereby the experimental group mean exceeds the control group mean. Our goal was to determine the relative frequency with which obtained p_{rep} statistics are close to true replication probabilities. Regardless of which way p_{rep} is defined, our simulations show that it is unlikely to be close to the true value unless both the population effect magnitude and the sample size are uncommonly large. The definitional problem in combination with the inaccuracy under either interpretation, constitutes an important challenge for those who espouse the routine computation of p_{rep} statistics.

Keywords: p_{rep}, replication, probability

There is much controversy surrounding a recent proposal by Killeen (2005a, 2005b, 2006; Sanabria & Killeen, 2007) that researchers compute the probability of replication (p_{rep}) rather than the traditional probability of the finding (or a more extreme finding) given the null hypothesis (p). To assess whether it would be desirable for researchers to compute p_{rep} routinely, it is important to determine the relative frequency with which obtained p_{rep} statistics are close to the true probability of replication. If the relative frequency is large, then that would support p_{rep}. But if the relative frequency is small, then there would be less reason for researchers to compute p_{rep}. Therefore, our main goal was to present computer simulations that test the relative frequency with which obtained p_{rep} statistics are close to the true value.

A secondary goal was to sort out the definitional issues concerning the p_{rep} statistic; that is, precisely what does p_{rep} mean? Such an understanding is a prerequisite for fulfilling our main goal; to perform the appropriate simulations, it is necessary to know how to define p_{rep} in the context of preset population parameters. We believe that much of the existing controversy over p_{rep} (e.g., Iverson, Lee, & Wagenmakers, 2009; Iverson, Lee, Zhang, & Wagenmakers, 2009) can be traced back to a basic definitional ambiguity.

Killeen's writings imply two definitions, and these two definitions are not compatible. Some criticisms against p_{rep} assume Definition 1 below (e.g., Iverson, Lee, & Wagenmakers, 2009; Iverson, Lee, et al., 2009), which is what we believe to be the popular definition, whereas Killeen's defenses against those criticisms tend to assume Definition 2. We address these in turn.

David Trafimow, Justin A. MacDonald, Stephen Rice, and Dennis L. Clason, Department of Psychology, New Mexico State University.

Correspondence concerning this article should be addressed to David Trafimow, Department of Psychology, New Mexico State University, MSC 3452, PO Box 30001, Las Cruces, NM 88003-8001. E-mail: dtrafimo@nmsu.edu

Killeen (2005a) defines *replication* as "[finding] an effect of the same sign as that found in the original experiment" (p. 346). In a later article, he states that p_{rep} "predicts replicability in general" (Killeen, 2005b, p. 1011). Thus, a reader of these articles might reasonably conclude Definition 1, which we believe is the usual interpretation, as well as the interpretation of most of the researchers who have criticized p_{rep}:

> *Definition 1:* p_{rep} is the probability of obtaining a result in Experiment 2 in the same direction as the result in Experiment 1, given the original result.

On the other hand, Killeen constructs a formal definition of the p_{rep} statistic in terms of observed effect size (d'; see Cohen, 1977): $d' = (M_E - M_C)/s_p$, where M_E and M_C are the experimental and control group means obtained during an experiment and s_p is the pooled within-group standard deviation. By this definition, positive effect sizes are observed when the experimental sample mean is greater than the control sample mean, and negative effect sizes are observed otherwise. p_{rep} is formally defined as $p(d'_2 > 0|d'_1)$ (see Killeen, 2005a, Figure 1; see also Killeen, 2005b). Definition 2 expresses this concept in words:[1]

> *Definition 2:* p_{rep} is the probability of obtaining a positive effect size in Experiment 2, given the original result.

It is important to note the contradiction between Definition 1 and Definition 2. Clearly, the probability of a result in the same direction as the original one (Definition 1) is not necessarily the same thing as the probability of a result in the positive direction, where "positive" means that the experimental group mean exceeds

[1] In agreement with ourselves, Doros and Geier (2005) similarly cited Figure 1 from Killeen (2005a) as indicating that p_{rep} is $p(d'_2 > 0|d'_1)$. In addition, they indicated that "Killeen's notation leaves room for varying interpretations of what $p(d'_2 > 0|d'_1)$ means" (p. 1005).

the control group mean (Definition 2). Whereas the Definition 1 interpretation of p_{rep} addresses the probability of replication, the Definition 2 interpretation of p_{rep} addresses the probability of finding that the experimental group mean exceeds the control group mean. Thus, according to Definition 1, p_{rep} can be argued to be well named because it addresses the probability of replication; but according to Definition 2, p_{rep} would be more appropriately named p_{pos} because it addresses the probability of obtaining a positive result in a future experiment.

There is a problem with Definition 1 that can easily be seen by imagining a typical experiment with a control condition and an experimental condition. Suppose that the true population effect (δ) is zero. The experimenter collects an initial set of data and observes an effect in a particular direction. Assuming that observed effect sizes are independent and identically distributed, and that the distribution of observed effect sizes is symmetric around its mean $\delta = 0$, the probability that the direction of the effect in a future experiment is the same as the observed effect is 0.5. That is, $p[\text{sgn}(d'_2) = \text{sgn}(d'_1)|\delta = 0; d'_1] = 0.5$. When estimating p_{rep}, the true effect size is typically unknown, so the observed effect size d'_1 is used as an estimate of the true effect size δ. The estimated probability of replication will therefore equal 0.5 only when the observed effect size (d'_1) is equal to zero, an event with effectively zero probability. For any nonzero observed effect size (i.e., for any experiment in which the means of the experimental and control groups differ), the estimated probability of replication must therefore be greater than 0.5. In other words, given an initial effect size in a particular direction, future results in the same direction are deemed more likely than future results in the opposite direction. Obviously, then, the p_{rep} statistic computed in any particular experiment will have a small probability of resulting in the correct value of 0.5 and a large probability of resulting in an overestimate of the true replication probability; p_{rep} will be upwardly biased because it is impossible to obtain p_{rep} less than 0.5.

The biased nature of the Definition 1 interpretation of the p_{rep} statistic can be even more problematic when nonzero true effects are considered (i.e., when $\delta \neq 0$). Consider another example where $\delta > 0$. In this case, positive effect sizes are more likely to be observed than negative ones. Imagine that an initial set of data is collected, d'_1 is calculated, and an estimate of p_{rep} is generated based on the value of d'_1. As before, the estimate of p_{rep} must be greater than or equal to 0.5. However, the true probability of replication depends on the direction of the effect observed in the initial experiment. If the initial observed effect is in the same direction as the true effect (in this example, if d'_1 is greater than zero), then the true probability of replication is greater than 0.5, and the estimate of p_{rep} at least has a chance of being reasonably accurate. However, if the direction of the observed effect is opposite that of the true effect (if $d'_1 < 0$), then the true probability of replication will be less than 0.5, and the estimates of p_{rep} will always be greater than or equal to 0.5.

The obvious defense against these problems is to deny the Definition 1 interpretation of p_{rep} and to assert instead the Definition 2 interpretation. But this defense also runs into problems. One problem is that Killeen (2005a) touts p_{rep} as having the important advantage of being easily interpretable as the probability of replication. But as we have already seen, the Definition 2 interpretation of p_{rep} addresses the probability that the experimental group mean exceeds the control group mean; it is not the probability of replication. A related problem can be illustrated with a simple example. Suppose that Researcher A and Researcher B perform precisely the same experiment, in which the population effect size equals zero. Researcher A finds that the experimental group mean exceeds the control group mean, whereas Researcher B obtains an effect of equal magnitude but in the opposite direction. Both researchers compute p_{rep}. If both researchers use the Definition 1 interpretation, they will each define the direction of their sample effect as positive, and both will obtain the same p_{rep} (greater than 0.5 in both cases). In contrast, if both researchers use the Definition 2 interpretation, Researcher A's result will be defined as positive, whereas Researcher B's result will be defined as negative, and so Researcher A will compute that p_{rep} exceeds 0.5, but Researcher B will compute that p_{rep} is less than 0.5. Thus, the conclusion will be that Researcher A has a greater probability of replication than Researcher B, even though both sample effect sizes were of the same magnitude!

One can circumvent this problem easily enough by asserting even more strongly that p_{rep} should not be interpreted as the probability of replication but rather as the probability that the experimental group mean will exceed the control group mean (i.e., the Definition 2 interpretation). But this assertion carries with it a steep price; it means abandoning the notion of a straightforward interpretation of p_{rep} as the probability of replication. The assertion also contradicts the quotations we provided earlier that imply that Killeen (2005a, 2005b) wants psychologists to use the Definition 1 interpretation. The appendix of the 2005a article further implies the Definition 1 interpretation by advising the researcher who obtains a negative finding to redefine the direction of the effect: "For consistency, if d' is less than 0, use $|d'|$ and report the result as the replicability of a negative effect" (p. 352). Clearly, using the absolute value of the obtained effect contradicts the Definition 2 assertion, brings us back to Definition 1, and causes Researcher A and Researcher B in the foregoing example to have the same p_{rep} (and one that exceeds 0.5); the p_{rep} aficionado cannot have both Definition 1 and Definition 2—he or she cannot have the cake and eat it too.

It is possible to argue for one other way of conceptualizing p_{rep}. Specifically, if the population effect size is known (or thought to be known), one can define *positive* as in the direction of the population effect, and interpret p_{rep} with respect to the population effect direction. But researchers who are interested in substantive issues usually attempt to investigate what is not known, rather than what already is known, and so there are likely to be few cases in which the population effect direction is confidently thought to be known. An exception might be when researchers consider it important to estimate the size of the population effect of a well-studied phenomenon. But in this case, why would the researcher be interested in p_{rep} statistics? Researchers who wish to estimate the population effect sizes of well-studied phenomena would use meta-analytic procedures to combine sample effect sizes across experiments to obtain a population effect size estimate; they would not use p_{rep} statistics.

At this point, we assume that the reader will agree that the p_{rep} statistic is conceptually difficult but perhaps it is at least accurate; perhaps the relative frequency with which p_{rep} statistics are close to true values is large. To help in assessing accuracy, let us define two types of p_{rep} statistics. According to the Definition 2 interpretation (see also Doros & Geier, 2005), p_{rep} is the probability of

obtaining a future positive effect, given an initial effect. Consistent with this interpretation, therefore, we define $p_{pos} = p(d'_2 > 0|d'_1)$. p_{pos} will be greater than 0.5 when the experimental group mean exceeds the control group mean and less than 0.5 otherwise. According to the Definition 1 interpretation, p_{rep} is the probability of obtaining a future effect in the same direction as the initial effect, given an initial effect. Consistent with this interpretation, therefore, we define

$$p_{same} = \begin{cases} p_{pos} & \text{if } d'_1 > 0 \\ 1 - p_{pos} & \text{if } d'_1 < 0. \end{cases}$$

A consequence of this definition, as we pointed out in our discussion of the Definition 1 interpretation of p_{rep}, is that p_{same} will always exceed 0.5. To keep straight which interpretation of p_{rep} we are using at any particular time, we use p_{same} and p_{pos} for the Definition 1 and Definition 2 p_{rep} interpretations, respectively.

Now that we have p_{same} and p_{pos}, how likely is either of these to be close to what they supposedly are measuring? Put more specifically, how likely is p_{same} to be close to the true replication probability, and how likely is p_{pos} to be close to the true probability of a positive result in the second experiment? Obviously, the interval designated to indicate closeness is somewhat arbitrary. We chose to define *close* as within ± 0.025 of the true value after noting that psychologists seem comfortable with 5% error rates ($\alpha = .05$ is a common significance level, and 95% confidence intervals are popular despite the fact that 5% of them do not include the parameter being estimated). We performed simulations to determine the proportion of p_{same} and p_{pos} statistics that are within this interval, at varying sample sizes and population effect sizes. Prior to our main simulations, however, we attempted to replicate the simulations performed by Iverson, Lee, and Wagenmakers (2009), using the p_{same} statistic to remain consistent with their interpretation of p_{rep}. One reason for this replication was to show that there is nothing idiosyncratic about our simulations; we expected to find that p_{same} is biased at low population effect magnitudes (consistent with p_{rep} critics) but that p_{pos} is less so (consistent with p_{rep} aficionados). A second reason was to test median p_{same} statistics as well as means, on the chance that using medians would result in less biased central tendencies. But we reiterate that our main goal pertains to the proportion of p_{same} and p_{pos} statistics that are close to true values; if there is little likelihood of being close to the true value, there is similarly little point in computing p_{same} and p_{pos} statistics, regardless of their bias or lack thereof.

Simulation Method

Overview

To understand the basic strategy, imagine that it was possible to perform an infinitely large set of experiments, each experiment containing control and experimental groups. Each experiment has an associated effect size, and therefore estimates of p_{same} and p_{pos} can be calculated. If the population effect magnitude (δ) is known, then the true values of these quantities are easily obtained. The mean or median p_{same} and p_{pos} scores could be compared with the true values to determine how much each measure tends to overestimate or underestimate the true value given that experimental and control conditions of size n are considered under each population effect magnitude. Secondly, criteria could be set for a desired accuracy interval, and the probability that the estimates are actually within that interval could be determined. Although it is impossible to perform an infinitely large set of experiments, it is possible to have a computer randomly generate sufficiently large sets of experiments so as to enable the analyses, at varying levels of n and varying population effect sizes.

Materials and Software

The experimental simulation was conducted with MATLAB (Version 2007b; MathWorks, Natick, MA) on an IBM-compatible computer with an Intel Q6700 CPU and 4 GB of RAM.

Procedure

Six sample sizes ($n = 10$, 30, 50, 100, 200, and 400 subjects in each of the control and experimental conditions) and 41 fixed population effect magnitudes ($\delta = -2.0$ to 2.0 in steps of 0.1) were used in the simulation. For each n–δ combination, 1,000,000 effect sizes (values of d'_1) were sampled from a normal distribution with mean δ and variance $2/n$ (see Iverson, Lee, et al., 2009, for details regarding the population parameters of the effect size distribution). This resulted in 246,000,000 experiments (6 sample sizes $\times$ 41 population effect sizes $\times$ 1,000,000 experiments). For each observed sample effect, the following statistics were calculated:

- p_{pos}. This is an estimate of the probability of a future positive effect, and we calculated it using Equation 6 in Killeen (2005a). The σ_{d_R} term in this equation is equal to $\sqrt{2}\sigma_d = 2/\sqrt{n}$.[2]
- p^*_{pos}. The true probability of a positive effect in a future experiment is equal to the area under the distribution of effect sizes to the right of zero. This equals $\phi(\delta/\sqrt{2/n})$, where ϕ is the standard normal cumulative distribution function.
- p_{same}. For positive values of d'_1, p_{same} is equal to p_{pos}. For negative values of d'_1, p_{same} is equal to $1 - p_{pos}$.[3]
- p^*_{same}. For positive values of d'_1, p^*_{same} is equal to p^*_{pos}. For negative values of d'_1, p^*_{same} is equal to $1 - p^*_{pos}$.

From these data, we were able to determine (a) the mean and median p_{pos} and p_{same} values across all experiments and (b) how often the estimates were close to the actual values.

[2] Whereas Killeen (2005a, Equation 3) uses an approximation to estimate σ_{d_R}, we see no reason not to use the actual value of σ_{d_R}, $2/\sqrt{n}$. Iverson, Lee, et al. (2009) discusses the derivation of the exact value for σ_{d_R} in more detail.

[3] It is trivial to show that this formula for p_{same} is equal to $\phi(|d'_1|/\sigma_{d_R})$, which is the formula given in the appendix of Killeen (2005a).

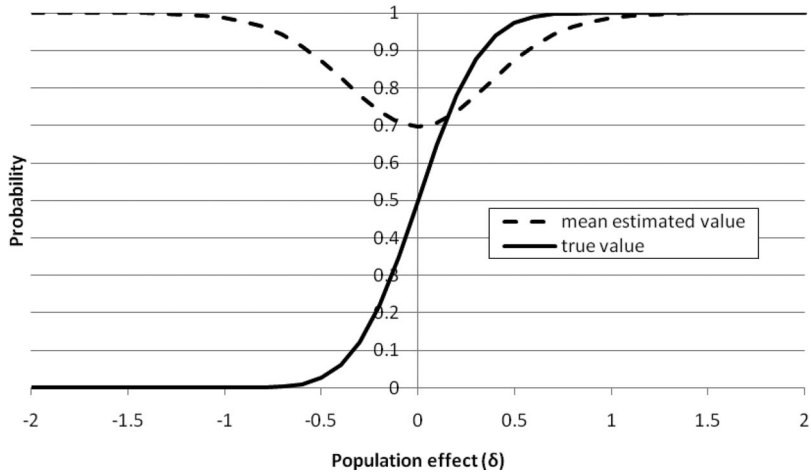

Figure 1. Mean estimated (p_{same}) and true (p^*_{same}) probability of replication as a function of δ when $n = 30$ and $d'_1 < 0$.

Results Concerning p_{same}

Comparing Mean and Median p_{same} Statistics With True Replication Probabilities

There are two possible values of p^*_{same} associated with any particular combination of δ and n, as the value of p^*_{same} depends on the sign of d'_1. When the initial effect is in the right direction (i.e., when $\delta d'_1 > 0$), p_{same} is likely to be a better estimate of p^*_{same} than when the initial effect is in the wrong direction. For this reason, we assessed the performance of the p_{same} statistic both when $\delta d'_1 > 0$ and when $\delta d'_1 < 0$. As expected, p_{same} diverged significantly from p^*_{same} when the direction of the observed effect was opposite that of the population effect. To illustrate, Figure 1 shows values of p^*_{same} and mean values of p_{same} when $n = 30$ and $d'_1 > 0$. In this example, p_{same} severely overestimates p^*_{same} for all negative population effects, a result that was replicated for all sample sizes used in the simulation. In general, the simulation results show that p_{same} is a large overestimate of p^*_{same} when $\delta d'_1 < 0$ regardless of sample size. In other words, if the experimenter is unlucky enough to observe an experiment effect in the wrong direction, then p_{same} will be a severe overestimate of the true probability of replication.

Of course, the probability of observing an effect in the wrong direction decreases as either $|\delta|$ or n increases, so one might argue that such extreme overestimates are unlikely for large sample sizes or population effect magnitudes. To avoid this criticism, Figure 2 depicts the true replication probability as the weighted average of the two possible values for p^*_{same}: $p^*_{pos} \times p(d'_1 > 0) + (1 - p^*_{pos}) \times p(d'_1 < 0)$. The figure also includes mean and median p_{same} statistics as a function of the population effect magnitude when $n = 10$, 30, and 50.[4] As a simulation check, note that in the case of mean p_{same} statistics, our results are similar to those obtained by Iverson, Lee, and Wagenmakers (2009), who used only means. Let us first consider Figure 2A, where $n = 10$. Figure 2A shows that the central tendencies of p_{same} statistics overestimate the actual replication probabilities when the population effect magnitudes in either the positive or negative direction are small (close to zero). However, at larger magnitudes, the p_{same} statistics quite accurately reflect actual replication probabilities. As the magnitudes increase even more, the curves diverge again such that the p_{rep} statistics underestimate the actual replication probabilities. Finally, as the magnitudes become extremely large, all the curves asymptote, and so mean and median p_{same} statistics accurately reflect true replication probabilities. It is further worth noting that because p_{same} statistics form a skewed distribution, median p_{same} statistics are sometimes slightly more accurate than mean p_{same} statistics, though these differences tend to be small.

It might seem strange that both mean and median p_{same} statistics overestimate the true value by such a substantial amount when it is 0.5. But as we discussed earlier, p_{same} will be greater than 0.5 for all nonzero observed effect sizes. Consequently, any reasonable central tendency measure (e.g., median or mean) of a set of p_{same} statistics, in which almost all the members of this set exceed 0.5, is going to result in a value that exceeds 0.5.

The results for larger sample sizes ($n = 30$ and $n = 50$ are shown in Figures 2B and 2C) were similar to those for $n = 10$ except that all curves are shifted and reach asymptote sooner. One positive consequence of the faster acceleration of the p_{same} curves as n increases is that the median and mean p_{same} statistics stop overestimating actual replication probabilities at lesser population effect magnitudes (in either the positive or negative direction). A negative consequence is that they also start overestimating actual replication probabilities at lesser magnitudes. In summary, p_{same} statistics sometimes overestimate and sometimes underestimate actual replication probabilities. One problem with having a bias that varies with the population effect magnitude is that it is more difficult to propose an easy fix to account for it, and so the fact that the bias is not constant can be considered to be a disadvantage.

[4] The full set of simulation results can be found online (http://psych.nmsu.edu/faculty/macdonald/personal/publications.html).

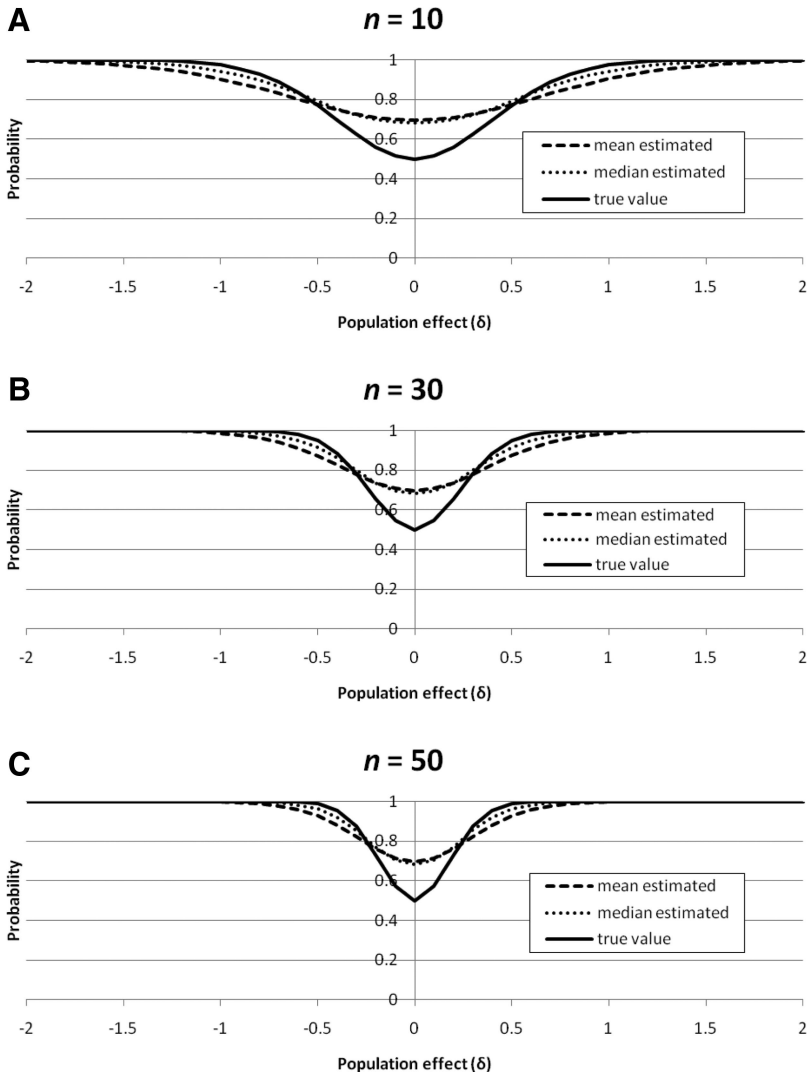

Figure 2. Represents the true probability of replication in same direction as initial finding (p^*_{same}), the mean p_{same}, and the median p_{same} as a function of effect size and n. Because the value of p^*_{same} depends on the sign of d'_1, the values of p^*_{same} depicted in the figure are weighted averages: $p^*_{pos} \times p(d'_1 > 0) + (1 - p^*_{pos}) \times p(d'_1 < 0)$.

How Often Are p_{same} Statistics Close to True Replication Probabilities?

We suggested earlier that although it is nice to know how well mean and median p_{same} statistics fit true replication probabilities, it is more important to determine the likelihood that calculated p_{same} statistics will be close to them. Figure 3 illustrates the proportion of trials for which the p_{same} statistics are close to the true replication probabilities (*close* is defined as within ± 0.025 of the true replication probability), for $n = 10$, 30, and 50. As Figure 3 indicates, at low population effect magnitudes (in either the positive or negative direction) and low sample sizes, p_{same} statistics are very rarely close to true replication probabilities. For example, when the population effect magnitude equals zero, p_{same} statistics are close to true replication probabilities 7.07%, 7.05%, and 7.03% of the time when $n = 10$, 30, and 50, respectively. When the population effect magnitude equals 0.2 (a "small" effect, according to Cohen, 1988), p_{same} statistics are close to actual replication probabilities 7.66%, 9.07%, and 10.66% of the time when $n = 10$, 30, and 50, respectively. When the population effect is 0.5 ("medium," according to Cohen, 1988), p_{same} statistics are close to actual replication probabilities 11.84%, 34.50%, and 44.59% of the time when $n = 10$, 30, and 50, respectively. Finally, when the population effect is 0.8 ("large," according to Cohen, 1988), p_{same} statistics are close to actual replication probabilities 26.98%, 63.68%, and 89.05% of the time when $n = 10$, 30, and 50, respectively. Although not shown in the figure, the trend toward better performance with increasing n continued out to $n = 400$. In general, however, at the effect magnitudes and sample sizes common in psychology, p_{same} statistics are close to true replication probabilities less often than might be considered desirable.

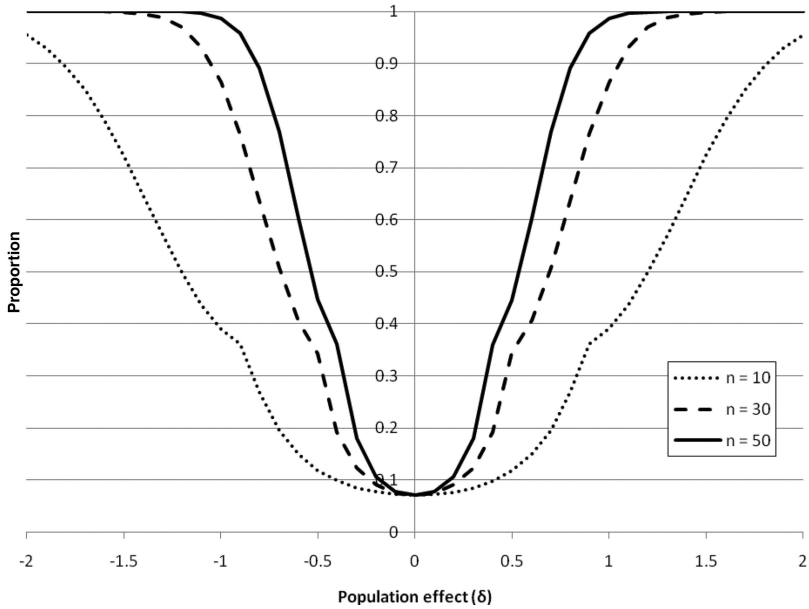

Figure 3. Proportion of p_{same} values that were within ± 0.025 of true p_{same} as a function of n. The x-axis represents effect size.

Results Concerning p_{pos}

There is an important difference between p_{same} and p_{pos}. In contrast to p_{same}, p_{pos} can take values less than 0.5. If a negative effect is observed ($d_1' < 0$), a future positive effect will be deemed less likely than a future negative effect, and therefore p_{pos} will be less than 0.5. The ability of p_{pos} to take values less than 0.5 will be reflected in the central tendencies (means and medians). However, when we consider the likelihood that any particular p_{pos} statistic will be close to the true value (p^*_{pos}), then results for negative population effect sizes will again be similar to those for positive ones.

Comparing Mean and Median p_{pos} Statistics With True Values

As Figure 4 shows, p_{pos} is less biased than p_{same} at low population effect magnitudes (in either the positive or negative direction); p_{pos} means and medians are reasonably close to true values. For example, when the population effect size is zero, and so the value of p^*_{pos} is 0.5, the means and medians tend to be very close to that value (within 0.01). It is only when the population effect magnitudes become reasonably large that p_{pos} becomes biased. For large positive population effect magnitudes, p_{pos} means and medians tend to be underestimates, but for large negative population effect magnitudes, p_{pos} means and medians tend to be overestimates. There are two qualifying factors. One of these is that as the sample size increases, bias starts to appear at smaller population effect magnitudes. Secondly, as bias increases, medians are noticeably less biased than means. For example, consider the case when the population effect size is 0.6 and the sample size is 10; the value for p^*_{pos} is 0.9101, the mean p_{pos} is 0.7808, and the median p_{pos} is 0.8282. Thus, the mean is biased by 0.0474 more than the median. Although the difference between means and medians is unimportant for the normal researcher who is interested in substantive issues and who calculates p_{pos} only one time for a particular experiment, the difference is important for the p_{rep} debate. Given that medians are less biased than means, and importantly so in some cases, using means can be interpreted to constitute a straw person criticism.

How Often Are p_{pos} Statistics Close to True Values?

Somewhat surprisingly, given a particular effect size (δ), sample size (n), and interval size, the probabilities associated with p_{pos} and p_{same} being close to their corresponding true values are equal, so Figure 3 for p_{same} also works for p_{pos}.[5] Obviously, then, as we have already seen that p_{same} often is not close to the true value under the population effect and sample sizes typical of psychology studies, a similar conclusion can be drawn with respect to p_{pos}.

Discussion

Figure 2 shows that mean and median p_{same} statistics overestimate true replication probabilities when the population effect sizes are small and underestimate them when the population effect sizes are large. In addition, although using medians rather than means improves matters at times, the extent of this improvement is barely noticeable. However, these simulations do not directly address the more important issue of how likely a calculated p_{same} statistic is to be close to the true replication probability. Figure 2 shows that under typical sample sizes and population effect sizes, the likelihood is not impressive.

Matters change a little if p_{pos} is used. The central tendencies are less biased in the conditions most common to researchers (e.g., low

[5] Details regarding this relationship are provided in the Appendix. Thanks to Geoff Iverson for providing the proof.

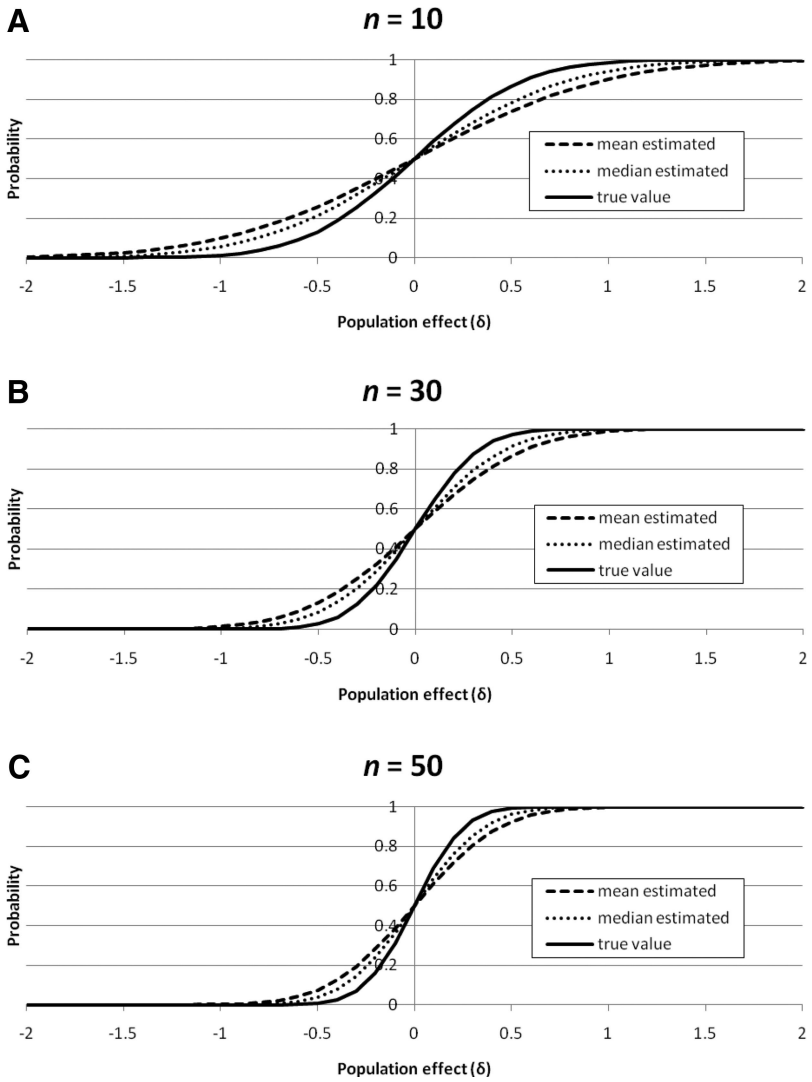

Figure 4. Represents the true probability of a positive effect (true p_{pos}), the mean p_{pos}, and the median p_{pos} as a function of effect size and n.

population effect sizes), which is a positive change from p_{same}. Also, in the cases in which p_{pos} is biased, the bias can be reduced noticeably by using medians rather than means, though we recognize that this difference might be important only for methodologists. The bad news, however, is that p_{same} or p_{pos}, as calculated for any particular experiment, is not particularly likely to be close to the true value unless the population effect sizes and sample sizes are larger than typical in the majority of empirical cases.

A potential limitation of the present simulations is that we based them on the population effect magnitudes and sample sizes, which resulted in complicated figures. An alternative would have been to combine these into a single parameter ($\Delta = \delta\sqrt{n}/2$), which would have resulted in more elegant figures (Iverson, Lee, et al., 2009). On the other hand, we believe that the reduction in elegance is more than compensated for by the fact that most researchers are used to thinking in terms of effect sizes and sample sizes but are not accustomed to thinking in terms of Δ.

Definitions Again

The simulations show that p_{rep} as interpreted via Definition 1 (p_{same}) is biased and unlikely to be close to the true value at common population effect sizes and sample sizes. The bias issue improves substantially when Definition 2 (p_{pos}) is used, but the likelihood of being close to the true value nevertheless remains problematic. Given that neither p_{same} nor p_{pos} is particularly likely to be close to the true value at the population effect magnitudes and sample sizes common in psychology, it seems useful to reconsider the underlying definitions, particularly as an explication of them was a prerequisite for performing the simulations. As we mentioned in the introduction, Definition 1 results in blatant bias at the conceptual level, even before running any simulations. And if this problem is avoided by resorting to Definition 2, then the topic has been switched; the probability of finding that the experimental

group mean exceeds the control group mean given an initial result is obviously very different from the probability of replication given an initial result. Thus, talking about "probability of replication" when what is meant is "probability that the experimental group mean exceeds the control group mean" is misleading.

The issue of two definitions prompts us to make recommendations for both those who would attack p_{rep} and those who wish to defend it. Much of the criticism has been based on using something akin to Definition 1, and if advocates of p_{rep} favor Definition 2, then this criticism takes on straw person characteristics. On the other hand, p_{rep} advocates should take responsibility for the fact that the verbal and mathematical definitions of p_{rep} are at variance—the former strongly implies Definition 1, whereas the latter implies Definition 2. The use of incompatible definitions increases the likelihood of misinterpretation and might be considered to be "asking for it."

Conclusion

Both definitions of p_{rep} cause serious problems. In the case of Definition 1, there is a strong bias at low population effect magnitudes and small effect sizes, and the bias is not lessened to any substantial degree by using medians rather than means. More important, the likelihood that the obtained statistic in any particular experiment is close to the true value is not impressive at the population effect magnitudes and sample sizes typical of psychology experiments. In addition, the probability of replication will be severely overestimated if the researcher is unfortunate enough to observe an effect in a direction opposite that of the true effect. Definition 2 suffers less from the bias problem, particularly if medians rather than means are used, but it suffers from the same low likelihood of being close to the true value that is a problem with Definition 1. Definition 2 also necessitates abandoning the notion that p_{rep} pertains to the probability of replication; this definition implies that p_{rep} pertains to the probability of obtaining a positive result. Therefore, we are unable to recommend that researchers compute p_{rep} statistics routinely. Perhaps we could make a more positive recommendation provided that either (a) advances are made to render them more accurate or (b) the conditions that simulations show to be necessary for accurate p_{rep} statistics are met. Of course, this recommendation depends on the definition of p_{rep} and the extent to which the final agreed-upon definition of p_{rep} actually addresses replication probabilities. We expect this issue to continue to be a problem.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

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

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

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., Lee, M. D., Zhang, S., & Wagenmakers, E.-J. (2009). p_{rep}: An agony in five fits. *Journal of Mathematical Psychology, 53,* 195–202.

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

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science, 16,* 1009–1012.

Killeen, P. R. (2006). The problem with Bayes. *Psychological Science, 17*(7), 643–644.

Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in the Schools, 44*(5), 471–481.

Appendix

p_{pos} Proof

For given values of δ and n, the probability that p_{same} will fall within the interval $(p^*_{same} - \varepsilon, p^*_{same} + \varepsilon)$ is equal to the probability that p_{pos} will fall within the interval $(p^*_{pos} - \varepsilon, p^*_{pos} + \varepsilon)$.

Proof:

$$
\begin{aligned}
p(|p_{same} - p^*_{same}| < \varepsilon | \delta, n) &= p(|p_{same} - p^*_{same}| < \varepsilon \cap d'_1 > 0 | \delta, n) + p(|p_{same} - p^*_{same}| < \varepsilon \cap d'_1 < 0 | \delta, n) \\
&= p(|p_{pos} - p^*_{pos}| < \varepsilon \cap d'_1 > 0 | \delta, n) + p(|1 - p_{pos} - 1 + p^*_{pos}| < \varepsilon \cap d'_1 < 0 | \delta, n) \\
&= (|p_{pos} - p^*_{pos}| < \varepsilon \cap d'_1 > 0 | \delta, n) + p(| - p_{pos} + p^*_{pos}| < \varepsilon \cap d'_1 < 0 | \delta, n) \\
&= (|p_{pos} - p^*_{pos}| < \varepsilon \cap d'_1 > 0 | \delta, n) + p(|p_{pos} - p^*_{pos}| < \varepsilon \cap d'_1 < 0 | \delta, n) \\
&= (|p_{pos} - p^*_{pos}| < \varepsilon | \delta, n)
\end{aligned}
$$

Received June 3, 2008
Revision received August 11, 2009
Accepted September 14, 2009 ■