Meta-Psychology, 2023, vol 7, MP.2022.3299
https://doi.org/10.15626/MP.2022.3299
Article type: Original article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Malte Elson
Reviewed by: František Bartoš, Eirik Strømland
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/VD4PN

How Can I Study from Below, that which Is Above? Comparing Replicability Estimated by *z*-curve to Real Large-Scale Replication Attempts

Lukas K. Sotola[1]
[1]Iowa State University

Z-curve is an analytic technique with which one can estimate the percent of a set of studies of interest that would replicate if one were to run actual replication studies. I compared the estimates *z*-curve yields to the outcome of real large-scale replication studies, such as the Open Science Collaboration (2015) work or the various Many Labs projects (e.g., Klein et al., 2014). I collected *p*-values from the original studies examined in six different large-scale replication efforts to the extent possible, ran *z*-curves on all the original studies, and compared the *z*-curve results to the results of the actual replication studies. My results show that across 163 replication studies taken from the six replication efforts, 85 (52.15%) showed statistically significant results in the expected direction as indicated by the authors of the replication studies. The outcome of the *z*-curve of all these studies was accurate, with the midpoint between the expected replication rate and the expected discovery rate, 50.55%, being almost exactly the same as the true replication rate. Its replicability estimate was also more accurate than that of *p*-curve analysis. Comparison of *z*-curve analysis of studies that did successfully replicate to studies that did not does suggest heterogeneity in the accuracy of its estimates, however. The pros and cons of *z*-curve analysis are discussed.

Keywords: *z*-curve, *p*-curve, replicability, replication studies

Replicability has become a central issue in psychological science (Nelson et al., 2018). While several events in the early 2010s gradually brought replication to researchers' attention, the event that is most widely cited as sparking the so-called replication crisis is a large study that attempted to replicate 100 effects in social and cognitive psychology and showed a replication rate of only 36% (Open Science Collaboration, 2015). Since then, primary replication studies have become much more popular: a PsycINFO search for replication studies yields 290 hits for the period between 2000 and 2011, while the same search for studies published after January 1st, 2011 yields 999 hits. However, it is not always practical to perform a large-scale replication effort, such as the Many Labs replication studies (e.g., Klein et al., 2014). Sometimes there are simply too many studies in an area to replicate them all, or one may want to assess the estimated replicability of a set of studies before expending a lot of time and effort and possibly money on a replication effort.

Consequently, analytic techniques for assessing replicability without performing replication studies have become popular as of late, both as tools in and of themselves and as supplements to meta-analyses. The most popular technique to this point is *p*-curve analysis (Simonsohn et al., 2014), but recently an extension of and improvement on *p*-curve called *z*-curve analysis has been created (Bartoš & Schimmack, 2020; Brunner & Schimmack, 2020). As *z*-curve is still relatively new and researchers may wonder about the accuracy of its estimates, the purpose of my current study is to assess the accuracy of *z*-curve's estimates of replicability. I do this by applying *z*-curve analysis to studies where actual replication attempts exist and comparing estimates of replicability yielded by *z*-curve to the replication estimates found in actual replication studies. I proceed by explaining to an extent how *z*-curve works as an analytic technique, including why it uses *p*-values as the unit of analysis in assessing replicability and how the technique itself calculates the estimates it yields. I then describe the need for a validation study on its accuracy and how I go about conducting such a study.

The Relationship Between *p*-values and Replicability

Z-curve analysis, like *p*-curve analysis, capitalizes on the relationship between the distribution of a group of *p*-values and replicability (and the presence of questionable research practices) in its analysis (Simonsohn et al., 2014). The fundamental insight that underlies the logic

of *z*-curve is that when researchers investigate true effects and conduct studies with strong statistical power (Cohen, 1992), then the *p*-values corresponding to the main statistical tests of such studies should tend to be disproportionately below $p = .01$. Further, *p*-values above $p = .01$ should be relatively rare, although not completely absent (Sellke et al., 2001). On the other hand, when a set of studies are under-powered and the investigators engage in questionable research practices, then there will tend to be a disproportionate number of *p*-values above .01 but below .05.

To illustrate why this is, I will discuss three hypothetical scenarios (Sotola, 2022). In Scenario I, a number of researchers study a phenomenon where there is no true effect—that is, the null hypothesis is true. For simplicity's sake, I will assume that the effect in question in all three of my scenarios is a mean difference between a treatment and control group. In this scenario, if the *p*-values yielded by the studies such researchers conducted were plotted as a histogram, they would form a flat plane, because all *p*-values would be equally likely to occur in the studies these researchers conducted. By extension, this means that 5% of the *p*-values these researchers find would be below .05. Assuming the investigators in question use the conventional threshold for statistical significance of $p = .05$ and assuming that journals tend to prefer publishing studies showing statistical significance (Rosenthal, 1979), they could get these 5% of studies published and argue based on them that they have discovered a true effect. However, those effects would only replicate at the same rate as the chosen threshold for determining statistical significance (=.05). Put differently, one would expect only 5% of the studies to replicate successfully.

In Scenario II, another group of researchers all investigate an effect where the true effect is small but not zero. Let us assume that the effect is around Cohen's d=.30. Let us further assume that these investigators are good about conducting studies with strong statistical power. Thus, they conduct all of their studies with sample sizes of 580, providing researchers with statistical power of around 95% at an alpha level of p=.05. Under these conditions, 95% of the *p*-values these researchers obtain will be below p=.05. However, crucially, a sample size of 580 also yields around 85% power when one uses an alpha level of p=.01, meaning that 85% of all *p*-values should fall below .01, while only 10% should fall between .01 and .05. If one were to plot these *p*-values as a histogram, one would find a right-skewed histogram, or one where most values were extremely low, and only a few were higher (Simonsohn et al., 2014). Moreover, one would expect that such studies would replicate rather well.

In Scenario III, I will assume the same base conditions as in Scenario II: an effect size of $d = .30$. But in this scenario, the researchers use sample sizes of 175, which provides only around 50% statistical power with an alpha of p=.05 and 27% power with an alpha of $p = .01$. This means that 50% of all *p*-values will be below .05, and only 27% will be below .01. Therefore, only around half of all *p*-values that are significant will be below .01, while the other half will be at various ranges between .01 and .05. The studies that show a statistically significant result will be the ones that are likely to get published, and so they are the ones that other researchers would likely be exposed to. But those would also be the studies most likely to have overestimated the true effect size (Gelman & Carlin, 2014), meaning that the effect sizes that get published will tend to be larger than d = .30. Accordingly, if other researchers attempt replication studies with the inflated effect size estimates in mind while conducting a priori power analyses, they are likely not to find that the previously published findings replicate at an ideal rate.

Comparing these three scenarios illustrates why one can make inferences about the replicability of studies based on the distribution of *p*-values. The studies yielded by Scenario I. are unlikely to replicate successfully simply because the underlying effect is null, and the published studies are based entirely on researcher degrees of freedom (Simmons et al., 2011). The studies yielded by Scenario II are likely to replicate, because the underlying effect is a true effect, and the original studies were conducted with strong statistical power. Finally, the studies yielded by Scenario III are likely to replicate at a slightly more favorable rate than those from Scenario I, but still are likely not to replicate at the same rate as in Scenario II. These three scenarios are of course extreme, and the reality in the field is likely some variation on all three depending on the size of the true effects being studied, whether there is a true effect, the practices of the researchers, and the practices of the journals to which the researchers in question tend to submit their work. They do, however, make it clear why the distribution of *p*-values can provide clues to the replicability of the studies in question.

Z-curve Analysis

This relationship between *p*-values and replicability is what *z*-curve capitalizes on (Bartoš & Schimmack, 2020). *z*-curve analysis is an extension of and improvement on the now-popular *p*-curve analysis. It uses the inferences one can make based on the distribution of *p*-values to estimate the replicability of the studies that generated the *p*-values entered into the analysis. This makes it similar to *p*-curve, except that in *p*-curve, *p*-

values are analyzed directly, whereas in *z*-curve, the *p*-values are converted to two-tailed z-statistics. Further, *z*-curve analysis assumes that each *p*-value comes from a different population of studies, whereas an assumption of *p*-curve analysis is that the studies that generated the *p*-values entered are from the same population of studies. Therefore, *z*-curve performs better under conditions of effect size heterogeneity (Brunner & Schimmack, 2020).

The unit of analysis in *z*-curve, then, is the aforementioned two-tailed z-statistics converted from the *p*-values extracted from the studies of interest. The z-statistics are actually the absolute values of the z-statistics, because one cannot know for certain the direction of the effect tested in the original study. Therefore, there are no negative z-statistics entered into a *z*-curve. Also, only significant *p*-values—or z-statistics above 1.96—are entered into the analysis, so the z-statistics analyzed are truncated at a value of 1.96. A finite mixture model estimated using these values is then estimated. Effectively, what this does is allow one to look at a set of studies as they appear after selection for significance and fill in the empty part of the distribution, or the complete distribution of z-statistics one would see if one had access to all studies conducted, published or unpublished. For those interested in learning more about the technical side of what *z*-curve does, they should seek out the original *z*-curve paper (Brunner & Schimmack, 2020), and I would highly recommend seeing Figure 1 in Bartoš and Schimmack (2020, p.9), as it makes what *z*-curve does more intuitive.

Once the complete distribution of z-statistics is estimated, one can assess the average power of the original studies (that were statistically significant) and the percent of the original studies that one would have expected to show statistical significance given the average power of the original studies (Brunner & Schimmack, 2020). The average power of the original studies is taken as an estimate of the percent of studies one expects to replicate if one were to conduct the studies in exactly the same way. This value is called the expected replication rate. The percent of studies that one would have expected to show statistical significance given the average power of all original studies—including hypothetically missing studies that may not be published—is called the expected discovery rate. The distinction between the estimated replication rate and the estimate discovery rate is that the former is the average power estimated for the original studies entered into the *z*-curve, while the latter is estimated based on the complete distribution, including hypothetically missing studies.

To my knowledge, *z*-curve has so far only been used a few times in the published literature (Bartoš & Schimmack, 2020; Schimmack, 2020; Sotola, 2022). This is understandable because it is relatively new compared to *p*-curve. However, some may remain skeptical of the accuracy of its estimates, specifically, about whether its estimates correspond to the actual percentage of studies that would replicate if one were hypothetically to run replication studies. I thought it would help to convince people of the validity of *z*-curve analysis to apply it to studies for which actual replication attempts exist. Moreover, it would be another mark in *z*-curve's favor if I could show that its estimates corresponded well to actual replication attempts.

To do this, I select several large-scale replication efforts—studies like the Open Science Collaboration (2015) and the Many Labs replication studies (e.g., Klein et al., 2014; Klein et al., 2018)—and apply *z*-curve to the original studies for which replication studies were conducted. This allows me to compare the estimated replication rate and estimate discovery rate the *z*-curves yield to the replication rate shown in the replication efforts as a way of assessing the accuracy of *z*-curve's estimates. I take several large-scale replication efforts, code the *p*-values from the original studies to the extent that that is possible based on the information the authors of the replication studies provide, and I apply *z*-curve to each individual replication effort for which I could code 10 or more *p*-values. I use 10 as a threshold, because the *z*-curve package in R does not allow the analysis to run if there are fewer than 10 significant *p*-values entered. Further, *z*-curve's estimates become extremely inaccurate with fewer *p*-values, and well more than 10 is usually recommended for running a *z*-curve analysis.

Doing this allows me to compare the estimated replication rate and estimated discovery rate to the outcome of each individual replication effort. In addition, I aggregate across all of the replication efforts from which I drew data (including those for which I was unable to run an individual *z*-curve), and perform a single overall *z*-curve of all of the studies included in all of the replication efforts I include. This allows me to do a single, overall comparison of the replication rate suggested by primary replication studies and the replication rate that *z*-curve predicts. This last analysis is the most reliable, because it includes the most *p*-values.

In addition, I make a few ancillary comparisons. First, I *z*-curve studies where the replication study was successful and studies where the replication study was unsuccessful separately. I do this to compare the performance of *z*-curve's estimates in the aggregate—across all replication studies—and its performance in narrower circumstances. Arguably, one of the weaknesses of *z*-curve analysis is that it is difficult to comment on the degree to which its results are meaningful for the out-

come of any one replication study. Applying *z*-curve analysis to studies where one already knows the outcome of an actual replication study may reveal underlying heterogeneity in the accuracy of *z*-curve's estimates that is masked by aggregating across all possible studies. The second ancillary comparison I make is between *z*-curve analysis and *p*-curve analysis. *z*-curve analysis builds upon *p*-curve analysis and is meant to be an improvement on it (Bartoš & Schimmack, 2020; Simonsohn et al., 2014). Indeed, past work has shown that *p*-curve overestimates the average statistical power of the studies included relative to *z*-curve analysis. This is because an underlying assumption of *p*-curve analysis is that all of the *p*-values entered come from effects that are all testing a single effect. In other words, *p*-curve assumes effect size homogeneity. However, *z*-curve analysis assumes heterogeneity: that each study could come from a different population of studies. To add to the comparisons others have made in this regard, I compare the outcome of a *p*-curve analysis to the outcome of a *z*-curve analysis, both for the overall analysis of all replication studies and for the comparison of successfully replicated to not successfully replicated studies.

I predicted that the estimated replication rate of each *z*-curve would be a slight overestimation of the actual replication rate, and that the estimated discovery rate would be a slight underestimation of the actual replication rate. As the creators of *z*-curve have pointed out, the estimated replication rate tends to be an overestimation of the percent of studies that are actually predicted to replicate (Bartoš & Schimmack, 2020). This is because the estimated replication rate is estimated under the assumption that the original studies can be conducted under exactly the same conditions as the original study was conducted. This assumption is rarely true, which means that the actual replication rate will end up being lower than the estimated replication rate due to both alterations in the conditions under which the replication studies are done and regression toward the mean. On the other hand, one can take the estimated discovery rate as a slight underestimation of the percent of studies one would expect to replicate. This is because the estimated discovery rate is the percent of studies—including studies that hypothetically may not have made it into published form due either to publication bias or QRPs—that one would originally have expected to show statistical significance. This is estimated with both the original studies entered into the *z*-curve and the hypothetical unpublished studies. Furthermore, I predicted that *z*-curve analysis will yield more favorable replicability estimates for the studies that successfully replicated as compared to those that did not; and I predicted that *z*-curve analysis's estimates would be more accurate than those yielded by *p*-curve analysis.

Method

Replication Study Selection

I decided to use widely cited and well-known replication efforts. These were: the Open Science Collaboration's (Open Science Collaboration, 2015) replication studies of 97 psychological studies; Camerer et al. (2018)'s replication studies of 21 social science studies published in the journals Nature and Science; Soto (2019)'s Life Outcomes of Personality Replication (LOOPR) Project, which ran replication studies for 78 correlations between the Big Five personality traits and significant life outcomes; and the Many Labs 1-3 projects (Ebersole et al., 2016; Klein et al., 2014; Klein et al., 2018). The Many Labs 1 project ran replication studies for 13 classic and contemporary psychological findings; Many Labs 2 ran replication studies for 28 classic and contemporary psychological findings; and Many Labs 3 ran replication studies for 10 psychological findings each. I did not include the Many Labs 4 study, because it only focused on replicating a single effect—the mortality salience effect (Klein et al., 2018). I also did not include the Many Labs 5 study (Ebersole et al., 2020), because it re-ran replication studies that originally had replication attempts included in the OSC's work, so I did not want to include some studies in the analyses twice. So my evaluation was based on six different replication efforts.

Data Extraction Method

In coding *p*-values to include in my *z*-curves, I relied exclusively on the exact test statistics from the original studies that the authors of the replication efforts reported either in the body of the paper they published or in supplemental materials. I used the test statistics I extracted to find exact *p*-values at the following link: Quick Statistics Calculators (socscistatistics.com). If I computed a precise *p*-value from a test statistic reported that was slightly above .05 but that the authors claimed was statistically significant, then I coded it as .049999 so that it would still be included in the *z*-curve when I ran the analysis. I assumed that if a replication study was being run on an effect, then it must have at least been treated as statistically significant in the original study, even if the exact *p*-value was slightly above .05. In addition to coding test statistics from each to-be-replicated study, I also coded whether the authors of the replication study indicated that the study successfully replicated. All the coding from my extraction process is posted on the Open Science Framework (OSF; https://osf.io/uge5r) page for this study. I now discuss

the coding process for each replication effort individually, because I had to use slightly different methods to extract data for my analyses from each replication study I selected.

Open Science Collaboration

Open Science Collaboration (2015) did not include test statistics or effect sizes or sample sizes from the original studies in their published paper. However, on their OSF page, they posted replication reports for each replication study that they completed and reported in their paper. These reports were written up by each separate team of researchers who ran each individual replication study. Many of them listed the test statistic corresponding to the finding from the original study which they hoped to replicate. This is where I extracted the test statistics to be included in my analyses. If there were multiple effects which the replication study was meant to replicate, I coded the one that they made clear was the crucial finding they hoped to replicate. If they did not make it clear which finding was the most crucial, I coded the first one that the authors of the report listed. In all, 66 of the 98 replication reports included enough information for me to extract test statistics, from which I could extract exact p-values. Near the end of each report, the authors summarized whether they believed they had or had not successfully replicated the finding in question, and this is how I coded whether the individual studies replicated.

Camerer et al. (2018)

Camerer et al. (2018) included effect sizes in the form of Pearson correlation coefficients and sample sizes from the original studies for which they ran replication studies in Table 3 of their supplementary materials. I used these correlation coefficients and sample sizes to extract exact p-values at the following link: https://www.socscistatistics.com/pvalues/pearsondistribution.aspx. Further to the right in the same table, there was a column indicating whether each study successfully replicated with a simple "yes" or "no," and I used this to code whether each study had successfully replicated. Thus, I was able to get p-values from all 21 studies for which Camerer et al. (2018) ran replication studies.

The Life Outcomes of Personality (LOOPR) Project

(Soto, 2019) included effect sizes and sample sizes from the original studies in Table 1 of the paper (p. 714). Most of these effect sizes were reported as Pearson correlation coefficients, but some were reported as standardized regression coefficients. I only coded those that were reported as Pearson's r, because it is not possible to compute an exact p-value using only a standardized regression coefficient, even if one has the sample size. Furthermore, there were cases in which the LOOPR Project ran replication studies of several effects that came from a single original study. In z-curve, as in p-curve (Simonsohn et al., 2014), the p-values one analyzes must be independent of one another. That is, they must be from separate samples. Therefore, I only coded the first effect size listed in the table from each study from which the LOOPR Project drew multiple effect sizes to replicate. It was not indicated specifically which effects the LOOPR Project replicated were from the same study in their table or their supplementary materials, but I assumed that if two or more effects in their table were listed alongside the same exact sample size that they were from the same original study. I computed exact p-values from the test statistics at the same link I used for Camerer et al. (2018)'s work. This extraction process resulted in 33 p-values I could analyze. I coded whether each study replicated using the "Replication success rate" column in Soto's Table 1. If the success rate was 100%, I coded it as successfully replicated, and if it was anything below 100%, I coded it as unsuccessful.

Many Labs 1

The Many Labs 1 effort did not list information from which I could compute p-values in their published paper (Klein et al., 2014). However, they posted a supplemental document in which they reported the original test statistics for six out of the 13 effects from the original studies for which they ran replication studies. I only included cases where the authors provided the test statistics reported in the original study reports, as that was the only way that I could find exact p-values from the original studies. I coded these six test statistics and extracted exact p-values using the calculators at the following link: https://www.socscistatistics.com/tests/. I coded whether the replication study was successful based on the overall result reported in the far-right column of Klein et al. (2014)'s Table 2 (p. 148). I note here that z-curve analysis does not run if fewer than 10 p-values are entered for analysis, so I could not run an individual analysis on the Many Labs 1 studies. However, I include these six studies in my overall z-curve analysis.

Many Labs 2

Many Labs 2 ran replication studies for 28 studies, and they listed test statistics from all the original studies in the text of their article, followed immediately by the outcome of their own replication studies (Klein et al.,

2018, ps. 453-467). I extracted the test statistics from the original studies and used them to compute exact *p*-values. I coded for whether the replication studies were successful based on what the authors wrote about the outcome of their replication study in the text immediately after they reported the test statistic from the original study. So for Many Labs 2, I was able to code all 28 studies for my analyses and had enough *p*-values to run an individual *z*-curve.

Many Labs 3

For the Many Labs 3 effort, I was able to extract *p*-values in the same way as for Many Labs 2, as the authors also reported test statistics from nine out of the 10 original studies for which they ran replication studies (Ebersole et al., 2016, ps. 73-77). I used these test statistics to find exact *p*-values, and coded for whether each study successfully replicated by looking at the null hypothesis significance tests reported in the far-right column of their Table 3 (p. 73). Thus, I was able to code nine *p*-values from this replication effort, meaning that I also could not run an individual *z*-curve on it. However, like with Many Labs 1, I include it in my overall *z*-curve analysis.

Analysis Plan

I ran individual *z*-curves on those replication efforts for which I was able to extract at least 10 *p*-values. Further, I aggregated across all replication efforts and ran a *z*-curve including all *p*-values extracted from all studies for which replication studies were run in one of the replication efforts from which I extracted data. I used the *z*-curve package in R (Bartoš & Schimmack, 2020) to run my analysis. While *z*-curve yields several estimates, I will focus on two: the expected replication rate (ERR) and the expected discovery rate (EDR). The former is effectively the average power of the studies included in the *z*-curve. The average power of the statistically significant studies is taken as an estimate for the percentage of the studies that one would expect to replicate if one were to do the studies again in exactly the same way. This is because statistical power is the probability that one will find a statistically significant effect if one studies a true effect (Cohen, 1992). The EDR is the percent of statistical tests that one have expected to be significant based on the average power of the original studies. The rest of the output from all of my *z*-curve analyses is posted on the OSF page for this project.

I focused on these two estimates, because they are the primary estimates of replicability that *z*-curve yields. ERR is a slight overestimation of the percent of studies one would expect to replicate, because it is computed under the assumption that each replication study will be done under the exact same conditions as the original study, a condition that is extremely difficult to achieve. Therefore, the actual percent of studies that would replicate should be slightly lower than the ERR. The EDR, on the other hand, is a slight underestimation of the percent of studies one would expect to replicate, because it is based on the average power of all studies (hypothetically) conducted, including those that did not show statistical significance. To compare *z*-curve's estimates to the results of actual replication studies, I compare the replication rates in the actual replication studies to the EDR and the ERR. I compute the difference between each *z*-curve estimate and the percent of studies that successfully replicated in reality to assess how close each estimate comes, and I also compute the midpoint between the EDR and the ERR. Presumably, the midpoint between the ERR and EDR should be the most accurate estimate of replicability, since the EDR is a slight underestimation and the ERR is a slight overestimation of the replication rate. I refer to the replication rates obtained in the replication efforts as the Actual Replication Rate (ARR) from now on for the sake of brevity.

Results

I have displayed all crucial results in Table 1. It shows the number of *p*-values I coded from each replication effort, the number of those *p*-values where the authors declared that their replication attempt was successful, the ARR, the ERR, the EDR, the difference between both *z*-curve estimates and the ARR, and the midpoint between the EDR and the ERR. I computed the differences between the EDR and ERR and the ARR by subtracting the ARR from the *z*-curve estimate. Thus, positive numbers indicate that *z*-curve overestimated the replication rate and negative numbers indicate that *z*-curve underestimated the replication rate. The *z*-curve of all *p*-values I coded is displayed in Figure 1.

Table 1

Comparison of Replication Rates in Actual Replication Studies to Estimates of z-curve

Replication Study	No. of *p*-values Extracted	No. Successfully Replicated	Percent Successfully Replicated (ARR)	ERR	EDR	ERRdiff	EDRdiff	ERR-EDR Midpoint
Soto (2019)	33	25	75.75%	94.30% [77.60%,100.00%]	75.10% [41.70%,100.00%]	18.55%	-0.65%	84.70%
Camerer et al. (2018)	21	12	57.14%	58.70% [29.20%,89.70%]	42.10% [5.00%,90.60%]	1.56%	-15.04%	50.40%
Klein et al. (2014)	6	4	66.67%	NA	NA			
Klein et al. (2018)	28	14	50.00%	46.20% [25.10%,75.40%]	13.60% [5.00%,64.80%]	-3.80%	-36.40%	29.90%
Ebersole et al. (2016)	9	2	22.22%	NA	NA			
Open Science Collaboration (2015)	66	28	42.42%	58.30% [39.00%,77.30%]	44.30% [7.00%,71.30%]	15.88%	1.88%	51.30%
All Combined	163	85	52.15%	62.20% [46.80%,73.80%]	37.90% [9.60%,65.30%]	10.05%	-14.25%	50.55%

Note. ERR = expected replication rate, EDR = expected discovery rate, ARR = actual replication rate, ERRdiff = the ARR subtracted from the ERR, EDRdiff = the ARR subtracted from the EDR, ERR-EDR Midpoint = the exact midpoint between the EDR and ERR.

As I expected, in all but one case, the ERR did, indeed, overestimate the ARR, and, again in all but one case, the EDR underestimated the ARR. In the four individual *z*-curves that I was able to run, at least one of the two *z*-curve estimates were within five percentage points of the ARR. The EDR was most accurate for Soto (2019, EDR = 74.10%, ARR = 75.75%), the ERR was most accurate for Camerer et al. (2018) (2015; ERR = 58.70%, ARR = 57.14%); the ERR was most accurate for Klein et al. (2018; ERR = 46.20%, ARR = 50.00%); and the EDR was most accurate for the Open Science Collaboration (2015, EDR = 44.30%, ARR = 42.42%).

While the outcomes for the *z*-curves of individual replication efforts are quite variable, both in terms of the confidence intervals they yielded and how accurate each estimate was, this is not surprising, as *z*-curve's estimates can be highly variable when few *p*-values are entered.

Thus, the overall *z*-curve analysis including all the studies I coded is the one to take most seriously. It shows that the overall ARR of 52.15% falls almost at the exact midpoint between the ERR of 62.20% and the EDR of 37.90%—50.55%. Put another way, *z*-curve analysis was only two percentage points off from the actual percent of studies that successfully replicated in real replication studies across 163 different replication studies.

Comparing Studies that Did Replicate to Those that Did Not Replicate

A *z*-curve on only the studies where the replication studies were successful (N = 86) revealed an estimated replication rate of 77.40% [60.00%,88.10%], an estimated discovery rate of 54.70% [16.20%,82.50%], a Sorić false discovery rate of 4.40% [1.10%,27.30%], and a file drawer ratio of 0.83 [0.21,5.19]. A *z*-curve on only studies where the replication studies were unsuccessful (N = 77) revealed an estimated replication rate of 37.90% [24.20%,57.80%], an estimated discovery rate of 22.20% [5.00%,44.50%], a Sorić false discovery rate of 18.50% [6.60%,100.00%], and a file drawer ratio of 3.51 [1.25,19.00].

Comparing *z*-curve Results to *p*-curve Analysis

A *p*-curve analysis run using the *p*-curve app (The *p*-curve app 4.06) with all of the studies coded yielded an estimated statistical power for the studies submitted of 99.00% [98.00%,99.00%], and the *p*-curve itself is shown in Figure 2. The half *p*-curve of only *p*-values below .025 suggested evidential value, z = -30.87, $p < .001$, as did the full *p*-curve, z = -32.15, $p < .001$; and the half *p*-curve suggested that statistical power was higher than 33%, z = 30.32, $p > .999$, as did the full *p*-curve, z = 21.75, $p > .999$. The *p*-curve of only studies where the replication studies were successful suggested evidential value both for the half *p*-curve, z = -30.06, $p < .001$, and the half *p*-curve, z = -31.7, $p < .0001$. The analysis yielded an estimated power of 99% [99.00%,99.00%]. The *p*-curve of studies where the replication studies were unsuccessful suggested evidential value for both the half *p*-curve, z = -12.85, $p < .001$, and the full *p*-curve, z = -13.69, $p < .001$. The analysis yielded an estimated power of 85.00% [78.00%,91.00%].

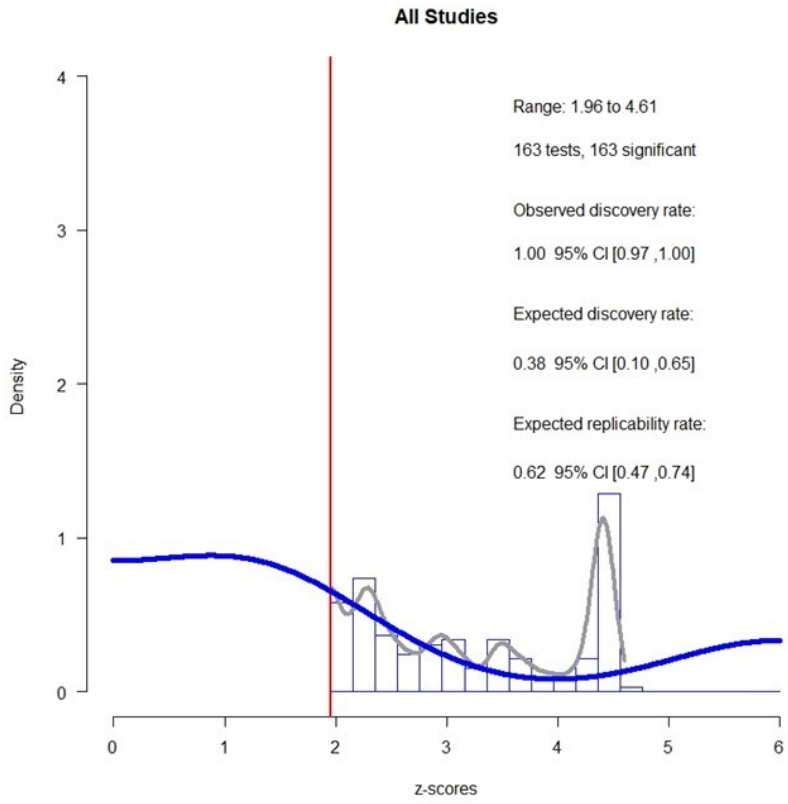
Discussion

My results suggest that, overall, *z*-curve analysis yields accurate replicability estimates, especially when one considers both the estimated replication rate and estimated discovery rate. This provides evidence for *z*-curve's real-world applicability: if one seeks to replicate a set of studies of interest, one can use *z*-curve to estimate the percent of those studies that will actually replicate if one were to run replication studies that mimicked the method and conditions of the original studies as closely as possible. *z*-curve's estimate of replicability is also much more accurate than the most immediately available alternative, *p*-curve analysis. My *p*-curve of the same set of test statistics yielded an estimated average power for the same set of studies of 99%, suggesting that 99% of them would successfully replicate. This is much more inaccurate than the estimated replication rate from *z*-curve analysis of 62.20%, and is in line with other work showing that *p*-curve tends to overestimate average power in the presence of effect size heterogeneity (Brunner & Schimmack, 2020). So this should encourage *z*-curve to be used more widely to assess the replicability of different areas of psychology.

That said, my comparison of only studies that successfully replicated and only studies that did not successfully replicate does reveal some variation in the accuracy of *z*-curve. Specifically, the accuracy of the *z*-curve of studies that did not successfully replicate and the *z*-curve of studies that did successfully replicate were not wholly desirable. Among the studies that did successfully replicate, the actual replication rate was 100%, and for the studies that did not successfully replicate, the actual replication rate was 0%, as indicated by actual replication studies. The corresponding *z*-curve analyses showed expected replication rates of 77.40% and 37.90%, respectively. While the trend of these two replicability estimates is in the right direction—the studies that successfully replicated showed a much higher estimated replication rate than the studies that did not successfully replicate—the estimates were, indeed, off by not unsubstantial margins. Therefore, while the

Figure 1

Plot of the z-curve Including All Coded p-values



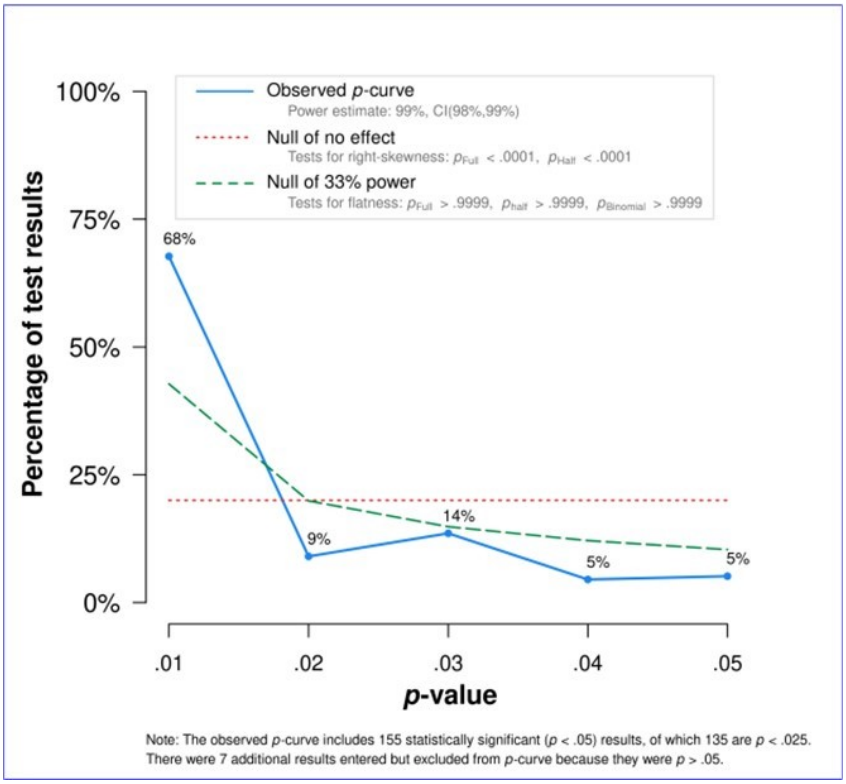
overall estimated replication rate was roughly accurate, this may mask heterogeneity in *z*-curve's accuracy.

This finding speaks to some reasonable reservations about *z*-curve's usefulness. While it may be effective at providing an aggregate assessment of replicability across a heterogeneous sample of studies, aggregate estimates of replicability may mask variation in the rate at which replication studies are successful and variation in the accuracy of *z*-curve's estimates across studies within different areas. I believe my comparison of *z*-curves of the studies that did and did not replicate hints at this, although a more comprehensive test of this is beyond the scope of my study. I will note here that past work testing the accuracy of *z*-curve's replicability estimates with simulation methods still assumed that a single effect size of interest was what was being tested (Brunner & Schimmack, 2020). The heterogeneity tested in that study regarded variation in a single main effect of interest, and not the heterogeneity that comes from many studies designed to test (in some cases) wildly different phenomena. I may tentatively suggest, then, that *z*-curve's replicability estimates may be most accurate when the studies reviewed are from the same area or test the same effect, even if there is heterogeneity within the studies (cf. Sotola, 2022). They may be less accurate if the studies reviewed are from a wide variety of areas (cf. Schimmack, 2020). If one does run a *z*-curve analysis of a wide variety of studies, then one should note the limitations inherent in such an analysis.

Moreover, there is still little to indicate that *z*-curve's

Figure 2

p-curve of All Studies Coded



replicability estimates are meaningful when one's focus is whether an individual study will replicate successfully or not. Because *z*-curve's replicability estimates are always aggregate, they do not say much about the replicability of any single effect one is interested in. If a researcher is interested in a single effect of a treatment, and a *z*-curve of studies on that treatment alongside unrelated studies is run, the estimated replication rate that *z*-curve yields does not necessarily speak to the replicability of the effect that that particular researcher is interested in. Possibly the only case in which *z*-curve's estimates might be meaningful in this way is if the studies included in the *z*-curve all test the exact same effect—for example, if studies included in the *z*-curve were only the studies from a meta-analysis of the treatment in question.

But even then, interpretation of replicability estimates would be complicated. If a researcher is focused on a single main effect of some treatment, and runs a *z*-curve on the effects from a meta-analysis of that treatment, and finds an estimated replication rate of, say, 40%, it might not be straightforward for that researcher to decide how to proceed. It is unclear if the 40% replication rate means that they should just proceed with their study but make sure the study is highly powered; whether they should run a replication study before running follow-up studies because the estimated replication rate is below 100%; or if they should simply abandon that area of research altogether due to low replicability estimates. Probably some combination of factors would

go into this hypothetical researcher's decision: the cost in both time and money of running a study on the effect; the real-world importance of the effect (e.g., one of theoretical importance vs. a treatment that might save lives); and the predicted size of the effect (cf. Anvari et al., 2022). A comprehensive discussion of these issues is beyond the scope of my work here, but it is an important issue that researchers should ponder.

Finally, I should also note that, when interpreting the outcome of a *z*-curve analysis, one should be sure to examine both the estimated replication rate and the estimated discovery rate. My results showed that in some of my individual *z*-curves, the estimated replication rate was more accurate, and in some of them, the estimated discovery rate was more accurate. So when evaluating the outcome of *z*-curve analyses, one should not fixate on either individually, but consider both, and, ideally, take note of the midpoint between the two values. In my overall *z*-curve, which should have yielded the most accurate results because it had the most *p*-values included, the actual percent of studies that replicated was almost the exact midpoint between the estimated replication rate and estimated discovery rate. Inasmuch as the former is an overestimation and the latter is an underestimation of the actual percentage of studies that are likely to replicate, I might tentatively claim that the estimated replication rate functions as an upper threshold for predicted replicability and the estimated discovery rate functions as a lower threshold for predicted replicability. Therefore, one should take the midpoint between them as the best estimate of the percent of studies that will actually replicate—an assertion which one of the creators of *z*-curve has endorsed (i.e., Schimmack, 2022).

The reservations I have pointed out notwithstanding, there is good reason to take *z*-curve's replicability estimates seriously (with caveats), even when one includes studies from a wide variety of areas in the analysis. The replicability analysis of only studies that successfully replicated and those that did not replicate also reveals that *z*-curve once again outperforms its most widely-known competitor—*p*-curve analysis. *p*-curve analysis yielded an estimated average power of 99% for the studies that did replicate, and 85% for studies that did not replicate. While the former is more accurate, the latter is much more inaccurate than the estimated replication rate yielded by *z*-curve analysis. I point this out here, because in assessing the merit of a new method, it is important not only to compare it to what would be ideal performance of the method—which one assumes in this case would be 100% accuracy of the estimated replication rate—but also to compare it to other available methods. If the new method outperforms the most readily available alternative, one can take that as a recommendation for the new method. *z*-curve seems to meet this criterion, at least when compared to *p*-curve analysis. Perhaps it is not perfect, but it seems to be superior to the most readily available and popular alternative for estimating replicability without running actual replication studies.

Conflict of Interest

I have no conflict of interest to declare.

Funding

I did not receive any funding to help with this project.

Author Contributions

The first author conceived of the project idea, conceived of the methodology, did all the coding, did all the analyses, and wrote and revised the manuscript.

Open Science Practices

This article earned the Open Data and the Open Materials badge for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Anvari, F., Kievit, R. A., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. C. (2022). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*, *18*, 503–507. https://doi.org/10.1177/17456916221091565

Bartoš, F., & Schimmack, U. (2020). *Z-curve.2.0: Estimating replication rates and discovery rates*. https://doi.org/10.31234/osf.io/urgtn

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, *4*. https://doi.org/1.15626/MP.2018.874

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behavior*, *2*, 637–644. https://doi.org/1.1038/s41562-018-0399-z

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., et al. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, J., R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, J., R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8. https://doi.org/10.1126/science.aac4716

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 364–376. https://doi.org/10.1037/cap0000246

Schimmack, U. (2022). *2022 replicability rankings of psychology journals*. https://replicationindex.com/2022/01/26/rr21/

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/https://doi-org.proxy.lnu.se/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/1.1037/a0033242

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Sotola, L. K. (2022). Garbage in, garbage out? evaluating the evidentiary value of published meta-analyses using z-curve analysis. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.32571