Original Manuscript

Social Psychological and Personality Science
1-9

DOI: 10.1177/1948550620938616
journals.sagepub.com/home/spp
SAGE

Type I Error Is Inflated in the Two-Phase Reverse Correlation Procedure

Jeremy Cone[1], Jazmin L. Brown-Iannuzzi[2], Ryan Lei[3], and Ron Dotsch[4]

Abstract

Mental images of social categories are highly consequential: They can reveal biases and help elucidate the factors that contribute to those biases. One strategy frequently used to evaluate the properties of mental images is reverse correlation, which is a data-driven method that allows researchers to visualize a person's mental representation of individuals or groups. In social psychology, this technique often employs a unique two-phase structure. This approach, however, has not yet been carefully validated, and its structure may alter the properties of the statistical tests used to evaluate differences between conditions. Using computer simulations to evaluate the Type I error rate in a typical two-phase reverse correlation procedure, we find that it is inflated in a nontrivial set of circumstances.

Keywords

reverse correlation, visual representation, Type I error, person perception, social cognition

Take a moment to conjure a mental image of a college professor. Is the person you called to mind young or old? Trustworthy or untrustworthy? Male or female? Mental representations of social categories such as this one are highly consequential: They can reveal biases and help elucidate the factors that contribute to them. At the same time, however, mental representations are difficult to study: They are complex, rich in detail, and their properties may not be readily available to conscious introspection or (honest) self-reporting. Indeed, some have suggested that mental representations are, in some ways, "ineffable" (Mangini & Biederman, 2004).

An influential technique for studying visual representations is called *reverse correlation*. This data-driven approach allows researchers to create a visualization of a person's mental representation (Brinkman et al., 2017; Mangini & Biederman, 2004) and, as a result, it has rapidly proliferated (for a review, see Brinkman et al., 2017, which cites over 25 papers that have used the procedure). However, despite its many advantages, there may be methodological issues with the procedure, at least as it is commonly employed in social psychology, as a result of its unique design employing two independent phases (described in more detail below). Here, we investigate whether the structure of a two-phase reverse correlation procedure may influence Type I error rates.

A Typical Two-Phase Reverse Correlation Paradigm

The procedure begins with an image (i.e., *base face*) onto which noise patterns are added that randomly distort its characteristics (Figure 1; see also Brinkman et al., 2017). On each trial of the *image generation phase* (Phase I), participants see pairs of images and their task is to select which of the two images more closely resembles their mental representation (e.g., "Which face looks more like a college professor?"). By making these judgments over many trials, participants are, in essence, providing information about the features of the images that correlate with their mental representation.

Selected noise patterns can be averaged together and overlaid on top of the base face to create a composite. Composites can be created both for individual participants (*individual CIs*) or, more commonly, for all participants in a condition (*group CIs*). Next, researchers conduct an *image rating phase* (Phase II) in which a separate group of raters judges the group composites (group CIs) on a target dimension of interest. Finally, a statistical test assesses differences in these subjective ratings (Figure 2).

Why Might Reverse Correlation Inflate Type I Error?

Unique to this procedure, group CIs created by one sample are evaluated independently by an entirely new sample in a

[1] Department of Psychology, Williams College, Williamstown, MA, USA
[2] University of Virginia, Charlottesville, VA, USA
[3] Haverford College, Haverford, PA, USA
[4] Anchormen Inc, The Netherlands

Corresponding Author:
Jeremy Cone, Department of Psychology, Williams College, Williamstown, MA 01267, USA.
Email: jdc2@williams.edu

separate image rating phase with its own methodology and design characteristics. Whatever variation existed in participants' responses in Phase I that created the group CIs is lost and supplanted by the variation in *raters' perceptions* of each group CI. In essence, the only information preserved from Phase I is the mean, but the standard deviation (*SD*) around the mean is lost in the transition to Phase II.

To see why this is problematic, imagine a hypothetical reverse correlation experiment in which we attempt to discover whether graduate (vs. undergraduate) students visualize college professors as warmer/friendlier. Unfortunately, however, we were only able to obtain two undergraduate and two graduate students to complete the reverse correlation procedure. (This is admittedly an extreme example for illustrative purposes, though to be clear, our simulations will explore experimental parameters more typical of actual reverse correlation experiments.) We then construct group-level composites for each of the two conditions. Because the sample size is so small, it is not difficult to imagine that some differences will emerge in the appearance of these group CIs even if there is no true difference in the two groups—perhaps even quite large differences. In a typical experiment, these differences would be considered in light of the standard error that emerges in each condition—that is, the sample size and variability that exists among the two participants in each cell.

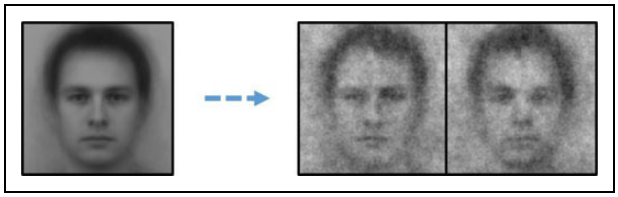

Figure 1. A base face (left) is superimposed with noise patterns that randomly distort its characteristics (right). Participants make judgments about which of the two randomly distorted images more closely resembles their mental image of a social category.

However, when we move to the image rating phase of a reverse correlation experiment, the sample size and variance that would normally be factored into the *t* test are lost and supplanted by the properties of the second phase of the experiment—that is, the number of *raters* and the variance in their *perceptions of the group CIs*. To take an extreme example, suppose that we decide to recruit 100,000 raters for each group CI (although again, to be clear, in our simulations we explore parameters more typical of actual experiments). Due to the large sample size, the standard error will be exceedingly small, and even very small differences in the ratings of the group CIs will be statistically significant. This could easily lead us to erroneously conclude that we have found evidence in favor of our hypothesis, even though the differences that emerged in Phase I could be entirely due to random chance based on the observations of an exceedingly small number of participants. This is the first problem with a two-phase procedure and it is explained in more detail in Figure 3 (top).

There is also a second problem with the structure of the procedure. When moving from Phase I to Phase II, an additional source of variation is introduced into the estimate of the effect size. In the first phase of the experiment, there is variance due to sampling error; if we were to repeat the first phase of the experiment, we would not expect to obtain identical group CIs every time. Instead, there is natural variation that will cause the group CIs to sometimes exhibit smaller differences and sometimes larger ones. The extent of this variation is captured by the sampling distribution in Phase I. Again, a typical independent-samples *t* test accounts for this variance by assessing whether the differences observed exceed those expected simply due to chance. In the reverse correlation procedure, however, these group CIs (that already exhibit natural variation) are then passed to the second group of raters who judge them on the trait of interest. However, these raters will *also* exhibit variance in their ratings of the group CIs. Even if we were to use identical group CIs in repeated runs of the experiment, we would

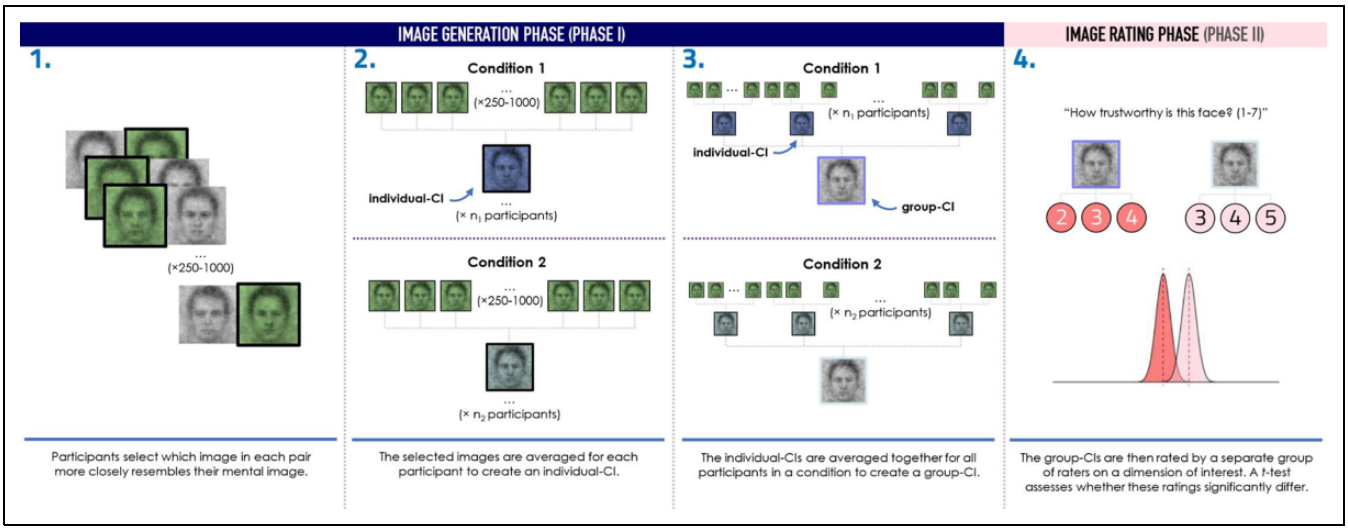

Figure 2. A two-image forced-choice reverse correlation procedure.

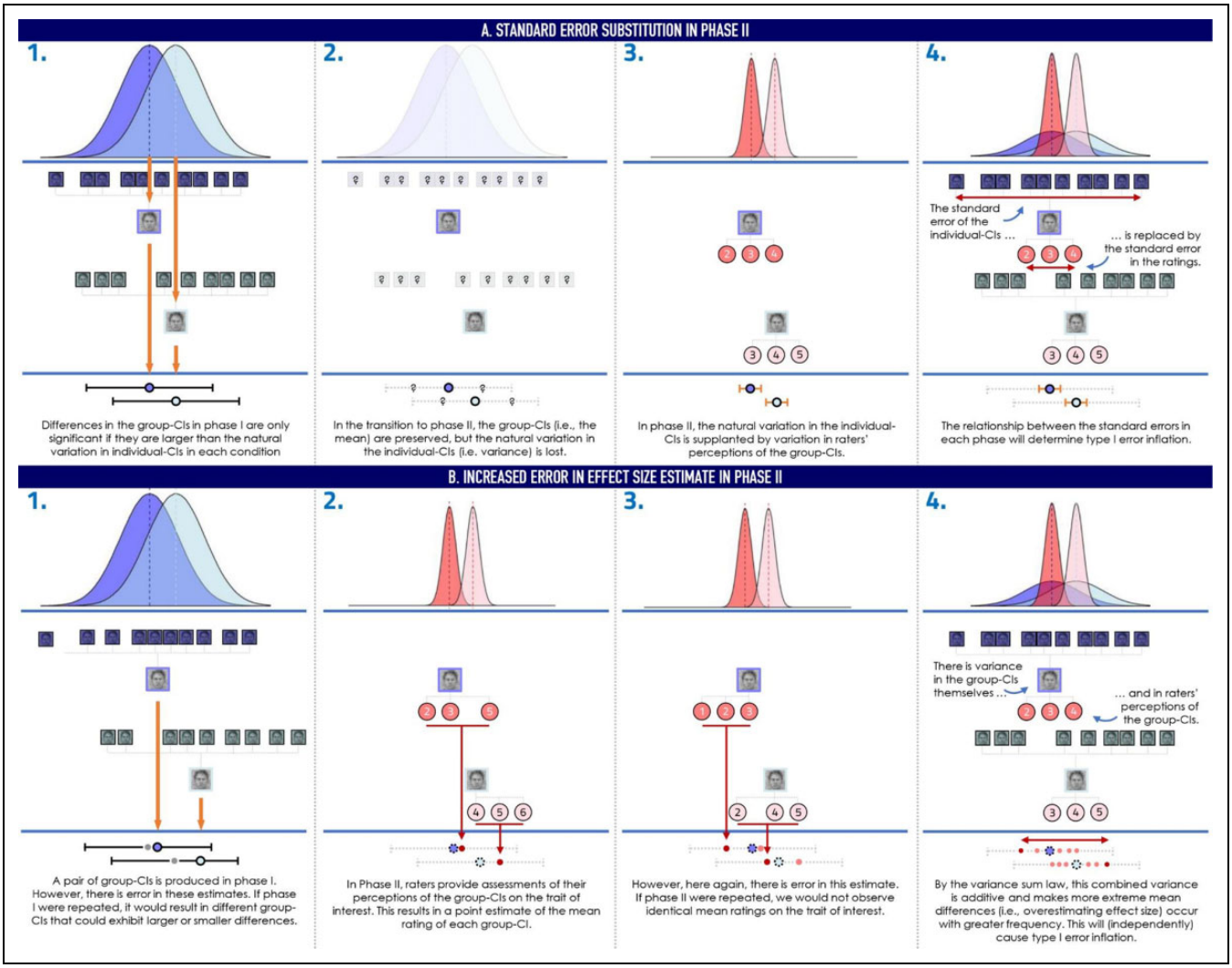

Figure 3. A summary of two reasons why a two-phase reverse correlation procedure could inflate Type I error. The first problem (top) is that the standard error in the image generation phase is substituted for the standard error of raters' perceptions of the group CIs. This will result in Type I error inflation to the extent that the standard error in Phase II is smaller than the standard error in Phase I. The second problem (bottom) is that, under repeated testing, the introduction of a second phase of the experiment adds additional sampling error into estimates of the effect size. There is error in both the first phase in the generation of the group CIs, and there is error in the rating of each pair of group CIs produced in Phase I. These two sources of variability are additive and make overestimation of the effect size occur with greater frequency.

observe different ratings on these group CIs with each run, as specified by the sampling distribution in Phase II. Importantly, this variation in the ratings is occurring independently of the variation in the group CIs that occurred in Phase I. This means that even if there are no true differences between conditions in Phase I, the random variation that occurred (again, as captured by the sampling distribution in Phase I) would mean that the group CIs are not identical when Phase II begins. The raters then exhibit additional variation in their ratings of these randomly varying group CIs. In sum, when combining the two phases of the experiment, under repeated testing, there are two factors that are independently varying: There is the variation that is occurring in the group CIs themselves under repeated testing, which causes them to exhibit different mean differences on each run of the experiment, and there is the variation that is occurring in the ratings task, which causes ratings to exhibit different mean differences on each run of the experiment.

This can also inflate Type I error. To see why, imagine, for example, that random variation happens to cause the group CIs to exhibit large differences but that these differences are not statistically significant when considered in light of the natural variation in Phase I. Nonetheless, in Phase II, if it happens that the ratings task *also* exhibits larger than average differences in the ratings, this increase could cause the *t* test to become statistically significant, even though it would not have been in Phase I. Indeed, the variance sum law specifies the exact variance of the combined distributions from both phases of the experiment—specifically, the two sources of variation are additive and, when combined, they ultimately widen the sampling

distribution of the differences of means in Phase II. The outcome is that extreme mean differences (i.e., overestimates of the effect size in the numerator of the *t* test in the ratings task) become more likely and make statistically significant results occur with greater frequency even under the null hypothesis.

This is analogous to the process that occurs in an independent-samples *t* test in which the sampling distribution of the differences of means has variance equal to the sum of the variances of the sampling distributions for each condition—that is, $\text{var}(M_1 - M_2) = \text{var}(M_1) + \text{var}(M_2)$. In a standard independent-samples *t* test, the mean of the sampling distribution of the differences of means is zero. However, the *variance* of the sampling distribution of the differences of means is much greater than that of a single-sample *t* test because of the variance sum law.

Importantly, this can occur despite the fact that, in the long run, the effect size estimate under the null hypothesis is zero (i.e., it is an unbiased estimator of effect size). In a two-phase reverse correlation procedure, under repeated testing, the effect size estimate in Phase II will randomly underestimate the effect size as often as it randomly overestimates it; these differences will cancel one another out in the long run. However, because there is increased *variance* in the estimate of the effect size (the numerator of the *t* test), more extreme differences become more likely to occur, just by chance, in the long run. This additional variance in the effect size estimate is not factored into the *t* test conducted in Phase II, which means that it will be too sensitive to differences that emerge in the ratings in Phase II—ratings that include *both* the variation in phase I *and* the variation in Phase II. This is the second problem with a two-phase procedure and it is explained in more detail in Figure 3 (bottom).

The more general observation is that assuming a two-tailed independent-samples *t* test with $\alpha = .05$, the statistical significance of the test is determined by two factors: (a) the effect size (i.e., the numerator of the *t* test) and (b) the standard error (i.e., the denominator of the *t* test). If, when moving from Phase I to Phase II of the experiment, both of these factors remain unchanged, then false positives ought to be successfully controlled. However, it is unclear whether either is preserved in practice: (a) the standard error between phases could be quite different depending on the sample size and natural variation that exists in each phase and (b) the estimate of the effect size between phases could be overestimated (in the long run) due to increased variability introduced as a result of resampling.

The Current Research

These observations motivated us to test Type I error inflation in the reverse correlation procedure. We conducted computer simulations that manipulated the sample size as well as the ratio of the *SD* of each phase of a simulated reverse correlation paradigm in which the null hypothesis was true (i.e., there were no systematic differences between conditions). We would expect $p < .05$ to occur at a nominal rate of $\alpha = .05$. However, over repeated iterations of this two-phase procedure, we can *directly observe* the probability of obtaining a statistically significant result. We predicted that the relationship between the standard errors of each phase would have a systematic relationship with Type I error rates.

Method

The computer simulations were conducted using the following six steps. The code and output for all of our simulations are available here: (https://osf.io/bvtzg/).

Generate a Set of Stimuli Using rcicr

We generated a set of stimuli using the rcicr package (Dotsch, 2016) using a neutral base image (a solid gray square). Typically, the base image used in actual research has a resolution of 128 × 128, 256 × 256, or 512 × 512. In our initial simulation, we used a resolution of 16 × 16 to keep computations manageable.

Create a Set of Phase I Participants

We simulated the image generation phase by randomly assigning *N* participants to two conditions of a simulated experiment. For each participant, we generated a set of random choices among each pair of images on each trial ($N = 300$).

Create Individual CIs

The randomly selected images for each participant were averaged using the standard reverse correlation procedure. The output is a two-dimensional square matrix for each participant.

Create a Group CI

The choices for all participants in each condition were averaged together to create group composites for each condition. The group CI can be thought of as a measure of central tendency (i.e., the "mean" of each condition).

Generate a Set of Phase II Ratings

In Phase II of an actual reverse correlation procedure, raters provide subjective assessments of the group CIs on a dimension of interest, such as whether a face appears masculine or feminine. We simulated these judgments by creating normal distributions that had means equal to the average brightness for each group CI and *SD*s set to a fixed value that varied in relation to the Phase I *SD* (across all iterations: $M = 0$, $SD = 0.0586$). This was determined by taking the average brightness of each of the individual CIs in each condition in Phase I and calculating the pooled *SD* of these values, just as we would if we were calculating an independent-samples *t* test on the individual CIs. Then, we multiplied this value by a multiplier to systematically vary the ratio of the *SD*s between phases, with values ranging from 0.001 to 10. For example, if the pooled *SD* in the image generation phase was equal to 5 and the multiplier was set to

0.5, then we would set the Phase II *SD* for both conditions to be $SD = 2.5$. Values of less than 1 for this multiplier indicate that the Phase II *SD* is smaller than Phase I, values greater than 1 indicate the opposite, and a value of 1 indicates that they are identical.

The outcome of the simulated Phase II is a set of N ratings drawn from two normal distributions ($N/2$ ratings each) with means equal to the values obtained on the group CIs and *SD*s that were systematically varied in relation to the Phase I *SD*.

Conduct a t Test on Simulated Ratings

Finally, we conducted an independent-samples t test to test for differences in the simulated image rating distributions. Because there are no true differences between the groups in our simulation, any significant differences are due to chance alone.

Steps 2–6 of this procedure were executed 10,000 times for each permutation of the simulation to give us a measure of the long-run probability of a false positive under each combination of manipulated variables. Our primary measure was the proportion of significant t tests that emerged over these 10,000 iterations.

Results

How Does the SD in Phase II (Relative to Phase I) Influence False Positives?

In our initial simulation, we fixed sample size in both phases to be $N = 100$ (50 per condition) and executed the simulation for different values of the *SD* multiplier (Figure 4). As the *SD* in Phase II becomes smaller in magnitude relative to Phase I *SD* (i.e., moving right to left in the figure), the likelihood of a false positive exponentially increases. These data were well fit by a three-parameter exponential decay function ($r^2 = .9993$).

A notable aspect of these results is that there is still inflation of Type I error beyond the theoretical value of $\alpha = .05$, even when the *SD*s between phases are identical (i.e., *SD* ratio = 1). This is because, although the *standard errors* are identical between phases (i.e., the denominator of the t test; Figure 3A), there is still additional error in the *estimate of the effect size* (i.e., mean differences in the numerator of the t test; Figure 3B). Indeed, due to this additional error, it isn't until the Phase II *SD* is *5 times larger* than the Phase I *SD* that the rate of false positives approaches the theoretical value of $\alpha = .05$.

How Does Sample Size Affect the Likelihood of a False Positive?

Our initial simulations held sample size constant to isolate the effects of relative differences in the *SD*s between phases. However, more important than the *SD* is the *standard error* in each phase of the experiment, which is simultaneously impacted by both the *SD* and the sample size. Thus, in the next set of simulations, we systematically manipulated both of these factors. First, we held Phase II sample size constant at $N = 100$ and varied Phase I sample size. Next, we did the reverse. (All sample size combinations for both phases are documented in Supplemental Material [SM].)

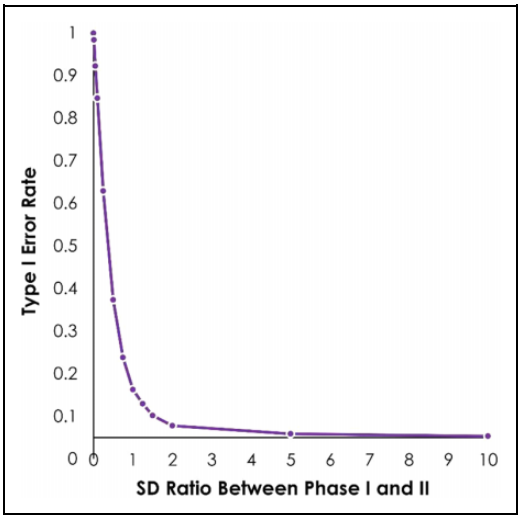

Figure 4. The false positive rate as a function of the relationship between the standard deviation (*SD*) in each phase of a simulated two-image forced-choice reverse correlation procedure. The sample size in both phases of the experiment for this iteration of the simulation was fixed to $N = 100$ (50 per cell). Each data point is the observed false positive rate across 10,000 simulated runs of the procedure. *SD* ratio values less than 1 indicate that Phase I has a larger *SD* than Phase II; values greater than 1 indicate that Phase I has a smaller *SD* than Phase II.

Phase I sample size. As the sample size in Phase I increases, the standard error decreases. Thus, mean differences between the group CIs in Phase I should also decrease, making it less likely that statistically significant differences will be detected in Phase II. This was confirmed by the simulation (Figure 5), which shows that as the sample size gets larger in Phase I, the likelihood of a false positive decreases (i.e., less area under the curve). However, even with a very large sample size in Phase I ($N = 500$), there is still inflation of false positives for many plausible values of the Phase II *SD* (including, as above, when Phase I and II have identical *SD*s). For example, if Phase I $N = 500$ and Phase II $N = 100$, the observed rate of false positives when the *SD*s are identical is still somewhat inflated at 0.075.

Phase II sample size. As the sample size in Phase II increases, the standard error decreases. Thus, smaller mean differences in Phase I will become statistically significant, increasing the likelihood of a false positive. Confirming this prediction, the simulations show that as the sample size in Phase II increases, the likelihood of a false positive increases (i.e., more area under the curve; Figure 6). However, even when Phase II sample size is relatively small ($N = 30$), there is still inflation of false positives for many plausible values of Phase II *SD*s. For example, if Phase I $N = 100$ and Phase II $N = 30$, the observed rate of false positives when the *SD*s are identical is still somewhat inflated at 0.081.

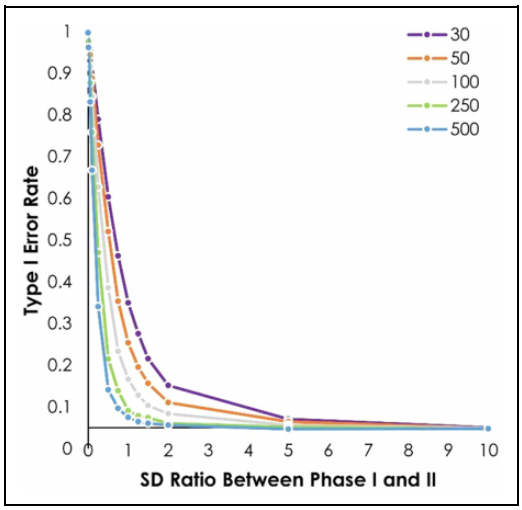

Figure 5. False positives as a function of manipulation of the Phase I sample size. Phase II sample size is held constant at $N = 100$ (50 per cell). As Phase I sample size increases, the likelihood of a false positive decreases, and smaller values of Phase II standard deviations are necessary for inflation to occur. Each observation is based on 10,000 iterations of the simulation.

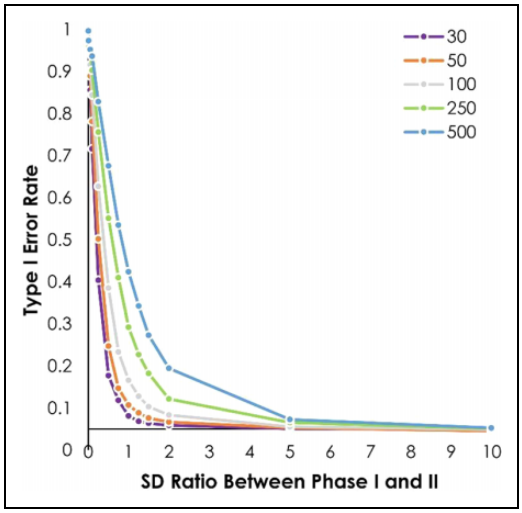

Figure 6. False positives as a function of manipulation of the Phase II sample size. As Phase II sample size increases, the likelihood of a false positive increases. Each observation is based on 10,000 iterations of the simulation.

How Does the Relative Magnitude of the Standard Error in Each Phase Impact False Positives?

These effects occur because standard errors are impacted by both the *SD* of the distributions and the sample sizes in each phase. However, if we reanalyze the data with a focus on the ratio of the standard *errors* between phases, we can assess the likelihood of a false positive using a method that is invariant to sample size. To accomplish this, we reanalyzed all of the permutations of the sample size manipulation simulations reported above ($N = 30$, 50, 100, 250, and 500 for each phase across 13 *SD* multiplier values). For each iteration of the simulation, we calculated the observed ratio of standard errors between phases. Then, we calculated the average of these values across the 10,000 iterations for each permutation ($N = 325$) and plotted them against the observed rate of false positives (Figure 7).

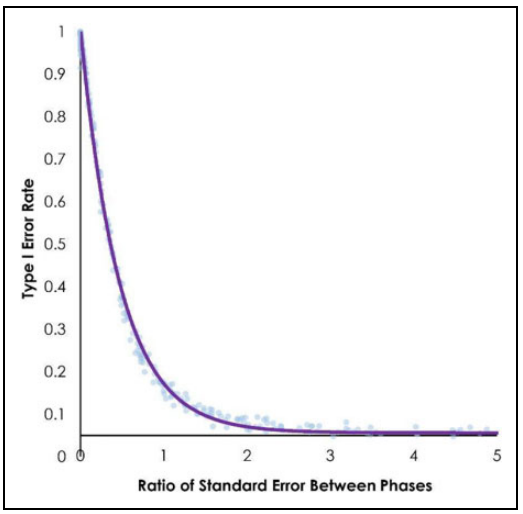

Figure 7. The observed relationship between the ratio of the standard errors between phases of the experiment and the observed likelihood of committing a Type I error. Each data point represents the observed ratio of standard errors (with values greater than 1 indicating that Phase I is larger than Phase II and values less than 1 indicating the reverse) and observed false positives over 10,000 iterations. The solid line depicts the predicted rate of false positives based on an exponential decay function ($r^2 = .9978$).

The data are, once again, well fit by a three-parameter exponential decay function ($r^2 = .9978$). These results suggest that, independent of the relative sample sizes in each phase of the experiment, Type I error is not sufficiently controlled until the standard error in Phase II is at least *3 times* the size of the standard error in Phase I.

Are These Results Robust to Changes in the Properties of the Procedure?

Finally, we sought to assess the robustness of these results to the manipulation of several aspects of the simulated procedure: (1) Phase II subjective rating measure, (2) number of Phase I trials, and (3) image resolution. These robustness checks show that the exponential curve we document in our initial simulation is robust to the manipulation of all of these factors, indicating that the patterns we observe are due to the general structure of the reverse correlation procedure rather than the specifics of how the task is carried out (see SM for more details).

Do Individual CIs Show a Similar Inflation of False Positives?

Another strategy used in the reverse correlation procedure is to rate the *individual* CIs on a subjective dimension of interest and

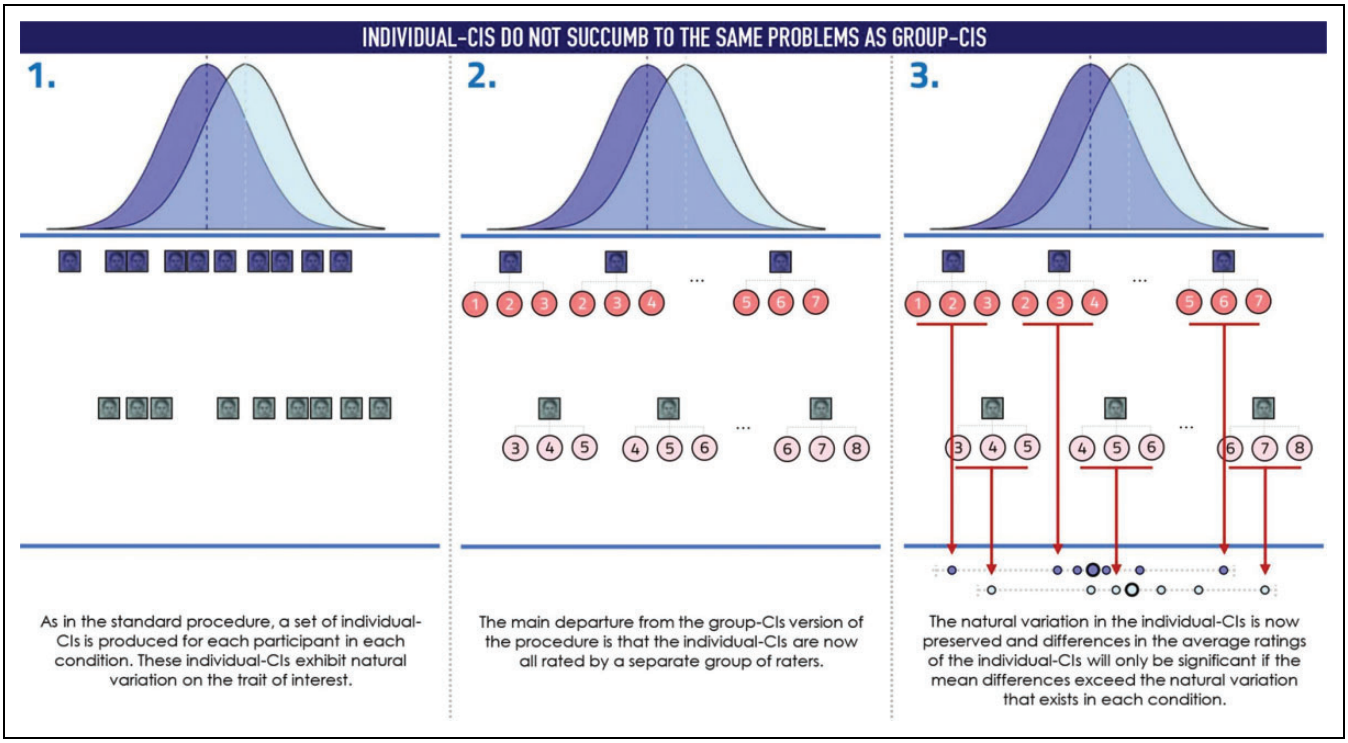

Figure 8. If the individual CIs are rated rather than the group CIs, the variance in Phase I is no longer lost and supplanted in Phase II.

to conduct a *t* test on these ratings (e.g., Dotsch et al., 2008; Young et al., 2013). This has the potential to avoid some of the pitfalls of group CIs in that the variance in the individual CIs in Phase I is preserved. Whereas for group CIs, the standard error in Phase II is determined by the number of raters and the variance in their perceptions of the group CIs, for individual CIs, the standard error is determined by the number of *individual CIs* and the amount of variability in the ratings of the composites themselves—that is, the individual CI approach retains the extent to which different individual CIs generate different ratings across all raters, rather than the extent to which people agree or disagree about the ratings of a single composite. Thus, the variance that occurs in people's mental representations among the participants in the image generation phase is preserved and presented to image raters in Phase II (Figure 8).

Are these methodological differences enough to control false positives? To test this, we conducted a variation of the simulation in which we modified Phase II to simulate a set of *N* raters who provided ratings for each of the individual CIs in Phase I. We used three possible values of the *SD* multiplier: 0.01, 1, and 10; however, because this manipulation had no effect on the rate of false positives, we report the average proportion of false positives collapsing across this variable. We used the mean rating over all raters as the measure for each individual CI. The values for all individual CIs in each condition were then submitted to an independent-samples *t* test (Table 1). In no circumstances was there an inflation of false positives beyond the theoretical value of $\alpha = .05$. Thus, false positives are successfully controlled when using individual CIs.

Table 1. False Positives as a Function of Sample Size in Both Phases of the Experiment When Conducting Statistical Tests on Individual CIs Rather Than Group CIs.

Phase I sample size	Phase II Sample Size				
	30	50	100	250	500
30	.048	.049	.050	.050	.051
50	.047	.047	.047	.048	.050
100	.048	.051	.049	.050	.052
250	.050	.052	.050	.049	.048
500	.050	.053	.050	.049	.050

General Discussion

Despite the many advantages of the reverse correlation procedure to the study of mental representations, our findings suggest that when using group CIs, there is a nontrivial set of circumstances in which Type I error is inflated beyond conventionally acceptable levels. However, Type I error is controlled when using individual CIs. Overall, our results suggest that for group CIs, the rate of false positives is a function of the relationship between the standard errors of each phase. Factors that decrease the standard error in Phase I (i.e., increases in Phase I sample size or decreases in Phase I *SD*) serve to increase the size of the ratio of the *SD*s and make false positives less likely. On the other hand, factors that increase the standard error in Phase II (i.e., increases in Phase II sample size or decreases in Phase II *SD*) serve to reduce the size of the ratio and make false positives less likely.

Assessing the Reliability of the Current Literature

The results of our simulations indicate that the published literature using reverse correlation procedures may be unreliable. Is there any way to evaluate the extent to which published findings may be unreplicable? Unfortunately, this is a difficult question to answer for at least three reasons. First, the Type I error rate was not inflated under every permutation of the simulations that we conducted. Second, determining the exact likelihood of a significant effect due to chance alone requires information about the sample sizes and *SD*s in *both* phases of the experiment—but in practice, we only know this information for Phase II. Third, and perhaps most importantly, even if the Type I error rate is theoretically inflated under certain permutations of the procedure, this merely indicates that the statistical significance of the test conducted on image ratings is inaccurate. Our simulations cannot shed light on whether or not a particular finding is "real" in the population.

These caveats aside, one strategy for assessing the potential magnitude of the problem is to explore the easily observable properties of past research—namely, sample size—to place some upper and lower bounds on the potential for Type I error inflation. The results of the simulation suggest that Type I error is especially likely to occur if the sample size in Phase II is larger than or equal to the sample size in Phase I (see SM for charts of all sample size combinations). When this is the case, the Phase II *SD* must be *at least* 2–3 times larger than Phase I in order for Type I error to be successfully controlled—an outcome that is unlikely to occur in actual rating tasks. However, a brief literature review ($N = 28$ studies reported in 24 published papers) suggested that 66% of past uses of reverse correlation had designs that had larger or equivalent sample sizes in Phase II relative to Phase I, meaning that a desirable ratio of the standard errors has likely not been consistently achieved in past published work.

What Can Be Done?

Our simulations suggest that the use of group CIs in a two-phase procedure is problematic. Can the procedure be modified to ensure adequate Type I error control? We propose three promising solutions.

Individual CIs. Our simulations suggest that there were no scenarios in which individual CIs caused inflation of Type I error. They are thus still a viable strategy for assessing condition effects in a reverse correlation procedure. However, two properties of individual CIs make them a less desirable or feasible option. First, their usage requires that dozens or perhaps hundreds of composites are rated, which can be practically infeasible. This places practical limits on the sample size of Phase I with important implications for the statistical power that can be achieved. Second, individual CIs are composed of many fewer individual forced-choice decisions. Whereas group CIs might be composed of thousands of judgments, individual CIs are generally composed of 300–1,000, thus having an undesirable influence on the signal-to-noise ratio. These two limitations are part of the reason why group CIs have proliferated so widely in the literature, and they suggest that it could be profitable to develop new techniques for reliably assessing differences in group CIs.

An objective metric. More objective comparisons of group CIs could reveal whether they exhibit sufficient differences that exceed what we should expect to occur by chance. Recently, Brinkman and colleagues (2020) proposed a promising technique for evaluating the informational value of a single composite. The logic of this test is very similar to a single-sample *t* test in which an observation is compared against a reference distribution to assess its likelihood of occurrence solely through the operation of a random process. This is accomplished by taking the composite—a square matrix of real numbers—and calculating a metric on this matrix called its norm. This observed value is then compared against a reference distribution in which composites are generated using a random process and the norm of each matrix is calculated. They find that this metric—which they call *infoVal*—can successfully distinguish between composites that are the products of systematic versus random responding. Theoretically, the logic of this test can be adapted to allow for a comparison of differences in the matrix norms between two group composites. This is a method that we are currently developing and testing.

A hybrid approach: Subgroup CIs. Calculating objective differences on group CIs is a useful strategy for assessing whether any true differences emerge between them, but ultimately, they cannot allow for tests of more specific hypotheses that predict the various *ways* in which the composites will differ on a subjective measure of interest such as trustworthiness or masculinity. Thus, there is value in developing an approach that preserves the subjective ratings phase of the experiment to test for more specific differences between them but that modifies the procedure in ways that prevent Type I error inflation. One possibility is to use subgroup CIs that are composed of the judgments of *random subsets* of multiple participants in each condition. Such an approach would have the benefit of preserving (some of) the variability in participants' visual representations from the first phase of the experiment (and thus potentially controling Type I error rates), while simultaneously (a) reducing the total number of images that must be rated and (b) increasing the total number of forced-choice decisions contributing to each subgroup CI. We have conducted an initial test of this approach and find that it does not inflate Type I error (Lei, Brown-Iannuzzi, Cone, & Dotsch, 2020).

Conclusion

These results indicate that Type I error is not sufficiently controlled in a typical two-phase reverse correlation procedure. Given these results, we suggest that researchers should not use group CIs as their sole strategy for assessing the effects of manipulation in a reverse correlation paradigm. Still, we should emphasize that Type I error

inflation is *not* inevitable. Individual CIs can still be used to test hypotheses until other methods for evaluating differences can be fully developed and validated. With an awareness of these pitfalls and the development of new strategies for assessing condition effects, reverse correlation still holds promise as a powerful technique for visualizing mental representations and making the "ineffable" amenable to empirical investigation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jeremy Cone https://orcid.org/0000-0001-5968-6711
Jazmin Brown-Iannuzzi https://orcid.org/0000-0002-2247-8385

Supplemental Material

The supplemental material is available in the online version of the article.

References

Brinkman, L., Goffin, S., van de Schoot, R., van Haren, N., & Dotsch, R. (2020). Quantifying the informational value of classification images. *Behavior Research Methods*, *28*, 333–361.

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*(1), 333–361.

Brown-Iannuzzi, J., Lei, R. F., Cone, J., & Dotsch, R. (2020). *Quantifying differences between group-CIs in the reverse correlation procedure* [Manuscript in preparation].

Dotsch, R. (2016). Rcicr: Reverse-correlation image-classification toolbox. *R package (Version 0.3)*, *4*. https://cran.r-project.org/web/packages/rcicr/index.html

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2013). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, *25*(2), 503–510.

Author Biographies

Jeremy Cone is an Assistant Professor at Williams College where he leads the Implicit Cognition and Evaluation (ICE) lab. His research concerns the factors that influence the formation and change of implicit evaluations.

Jazmin L. Brown-Iannuzzi is an Assistant Professor at University of Virginia. Her research seeks to understand why social group disparities may persist and the consequences of these disparities.

Ryan Lei is an Assistant Professor of Developmental Psychology at Haverford College. His research takes an intersectional framework to investigate how children acquire and apply stereotypes and prejudices.

Ron Dotsch is Senior Data Scientist at Anchormen working on AI applications. Academically, his research focuses on social face perception and the reverse correlation method.

Handling Editor: Margo Monteith