# Common statistical practices in IRAP research produce very high False Positives Rates: A simulation study

# Ian Hussey

# Common statistical practices in IRAP research produce very high False Positives Rates: A simulation study

# Old

## Description of simulations

In each simulation, data for each simulated IRAP trial type were randomly generated using a known population distribution (e.g., a true null effect represented by a normal distribution with M = 0 and SD = 1). This was done a large number of times (1000 for all simulations) in order to observe the long run error rate. A repeated measures ANOVA was then applied to this data, and then the significance of each effect was then assessed (i.e., whether *p* < .05 for the main effect for domain, main effect for trial type, and their interaction). Given that the simulations (unless noted otherwise) drew data from true null effect, all significant results were therefore false positives by definition. Such null results are generated because *p* values follow a flat distribution under the null hypothesis. That is, under the null hypothesis, all possible values (from 0 to 1) are equally likely to be observed. The acceptable false positive rate (FPR) is defined by the alpha value used as a cutoff for *p* values. The accepted norm for this cutoff of 0.05 implies that a 5% False Positive Rate (FPR) is deemed to be an acceptable long run false positive rate. Simulations that demonstrate a FPR that is higher than this value therefore imply that, under the conditions captured by the simulation, researchers using this analytic approach would generate false conclusions at this higher rate.

## Hidden multiplicity in ANOVA designs

Craemer et al (REF) demonstrated that although ANOVAs are extremely common in psychological research, they are subject to an false positive rate that is both inflated and yet underappreciated by many researchers. Put simply, imagine that a researcher is conducting a simple 2X2 between subjects experiment, and that she is willing to report and interpret results from any of the three *p* values that it produces (main effect A, main effect B, interaction effect A\*B). An alpha value of 0.05 is used, so each of the three significance tests has a 0.05 probability of producing a significant result under the null hypothesis. Because three tests are being conducted, the probability of observing at least one (False Positive) significant result is now 1-(1-alpha)^3 = 0.14, far higher than the 0.05 probability that might be incorrectly assumed from the alpha value.

## FPR for 4X2 mixed within-between ANOVA

While the FPR of simple designs such as this can be worked out purely mathematically, it is more efficient to quantify the FPR of more complex designs such as those used in IRAP research via simulation. However, it should be noted that the results of the two approaches are asymptotically identically (Craemers et al., REF). All analyses were done in R, and the scripts to implement them are available on the OSF (XXX).

Simulation 1 extended the above example used in Craemers et al from a 2X2 design to the 4X2 ANOVA design typically employed in IRAP research (e.g., 4 trial types and 2 between participants groups). Each simulated IRAP trial type was sampled from a normal distribution with M = 0 and SD = 1 (i.e., a true null effect). The four trial types were generated from independent correlations on the basis that, although analyses of IRAP data typically employs repeated measures ANOVAs, no correlation is actually assumed between the trial types. Indeed, the choice to employ RM ANOVA over the likely more appropriate MANOVA is one of historic convenience on the basis that the two generate similar results under most circumstances (REF). Group was simulated by alternating group assignment sequentially between the simulated participants (i.e., odd numbered participants assigned to group A, even numbered participants assigned to group B). All simulations reported here simulated data for 100,000 participants to estimate the long run false positive rate that would be encountered across highly powered experiments. FPR was observed to be 0.17. ….. Simulation 2 changed the sample size to twice the median sample size per cell employed in published IRAP research (median = XX; twice this given that we simulate a two groups design). This value was observed in the systematic review reported in section XX of this paper, below. This simulation therefore aimed to simulate the long run FPR that would likely be observed under real modal research conditions. FPR was observed to be 0.137.

The review of IRAP research practices (reported in section XX) observed that more complex designs have also often been used, such as 4X2X2 (4 trial types, 2 block order groups, 2 sample groups such as depressed vs. non depressed individuals). Simulation 2 extended the factorial design to 4X2X2 and observed a FPR of XXX. This illustrates the general rule that increasing the number of hypothesis tests (e.g., to three main effects, 3 two-way interaction effects and 1 three-way interaction effect) increases the FPR.

The review of IRAP research practices also highlighted that a variety of factorial designs have been used and analysed, sometimes by the same authors between studies or articles. This includes articles on which authors of the current article were coauthors. This suggests, and is our experience, that some procedural manipulations (e.g., block order) have been differentially analyzed or reported, sometimes on the basis of the significance of the results obtained (e.g., hussey REF). This “experimenter degree of freedom” is now more widely recognized as a form of *p* hacking that serves to increase the false positive rate. This inflation was estimated in Simulation 3 by, in each iteration, generating data and testing both 4X2 and 4X2X2 ANOVAs to it, and coding results as a false positive should either model generate a significant result. Simulation 3 produced a FPR of XXX. That is, when the true effect is null, nearly one third of IRAP studies employing this analytic strategy that has been employed in the published literature will generate false positive results.

The review of published IRAP research, along with our own direct experience of publishing such research, revealed that IRAP studies also differentially process and analyse IRAP data in another way. The IRAP natively produces four trial types (e.g., XXX), which can be entered as separate variables in the (modally used) RM ANOVA. However, these trial types can also be collapsed in a variety of ways: for example, data can be reduced to two trial types in one of two ways (e.g, concepts … attributes…), or to a single trial type that averages all four trial types. Simulation 5 generated data and applied all four data processing strategies (4 trial types, 2 concept trial types, 2 attribute trial types, 1 overall D score), fit ANOVAs to each, and returned the false positive rate across all models. An FPR of XX was now found.

At least two additional experimenter degrees of freedom are common within the IRAP literature. First, how participants are excluded from analyses on the basis of their performance with the IRAP. At least XX methods have been used across papers, sometimes different strategies being employed by the same authors. Second, whether one or more trial types are reverse scored and then analysed. This practice is routine but not standardized, and has been applied in a variety of ways across papers. Should these strategies not be defined ahead of time (or failing that, directly replicated), they likely serve to inflate the FPR further. However, these possibilities will not be simulated exhaustively – the take home point is that experimenter degrees of freedom that can be observed in the literature serve to inflate the FPR to worryingly high levels.

Simulation XX examined the impact of the generic IRAP trial type effect on the FPR.

In summary, these simulation studies simply instantiate the same assumptions made by the modal IRAP data analysis strategy of RM ANOVA applied to D scores, and observe the long run false positive rate caused by 1) the hidden multiplicity in ANOVA designs that goes widely unacknowleged and uncorrected across psychology research, 2) the known experimenter degrees of freedom in IRAP research, and 3) the observed generic IRAP effect. Whereas the typical alpha value of 0.05 suggests a 5% FPR, hidden multiplicity in ANOVAs increases the FPR for typical IRAP design to roughly 14%. Degrees of experimenter freedom can greatly increase the FPR but this depends on their severity, ranging from XX to XX. Finally, the known generic IRAP effect across trial types produces domain false positives in 100% cases in the long run of well powered studies. That is, where a significant p values gives rise to the (sometimes implicit) inference that participants demonstrated a domain specific bias (e.g., demonstrate a white people positive bias rather than show a significant effect between trial types that would also be shown for any non-racial stimuli included in the task). This suggests that most published IRAP research contains at least some false conclusions.

## Recommendations

Implications for the interpretation of existing work:

Published IRAP work should be read and interpreted with great skepticism. In particular

Ways to lower the FPR:

See cramer!

Alpha corrections for multiple testing.

Lower your alpha/justify your alpha.

Ignore main effect for trial type, focus on within-trialtype-and-between-conditions differences or criterion associations.

Bear in mind that this doesn’t help control FPR inflation from other sources, including EDoF/p-hacking. Robustness tests, comprehensive reporting via preregistration, open code and open data, can improve these.

# New

## Intro/why should I care bits

False-positives can be costly in scientific research, as they represent situations in which researchers and stakeholders incorrectly infer that a hypothesis is correct. For example, by mistakenly treating a hypothesis as correct they may dedicate future resources into that line of thinking.

…

Solutions to inflated FPR when using ANOVAs are known within the statistics literature, and references to them are common in statistics courses in psychology degrees. These include omnibus ANOVA tests and family-wise error corrections such as the Bonferroni–Holm correction. However, these solutions are rarely applied in practice. A systematic review of the use of multi-way ANOVA in 6 leading psychology journals revealed that this type of analysis is very common (present in 48% of articles), but the use of method to control the False Positive Rate are very rare (only 1% of articles; Cramers et al., 2016 REF). These results suggest that, at least up until recently, researchers either (1) may not be aware of the degree inflation in the False Positive Rate associated with these statistical methods; or (2) they may not be aware of the implications of this inflation; or (3) they are aware of both the inflation and its impact quality of our inferences, but are not concerned by it. Relatively less can be done about the latter reason given that it represents abdication of our scientific principles, but the former two reasons can be improved through increased training and resources. This article is intended to contribute to this effort to improve our collective future research.

…

One advantage of simulations is that you can simulate a wide range of conditions in order to understand how different factors influence the False Positive Rate.

## What is a simulation study?

These simulations employ the same assumptions used in experiments with actual participants. Random number generators are used to simulate samples of participants being drawn from populations with known parameters. Common statistical inference tests are then applied to the simulated data in the same way as they would be to real data. Inferences about the population effect (i.e., via the significance of *p* values) are then checked against the known population values used to generate the data. For readers who are unfamiliar with simulation studies, the following sections expand on their logic and provides a primer on simulation studies within the context of Null Hypothesis Significance Testing. All simulations in this manuscript and their results can be reproduced using the code available on the OSF (REF).

### Null hypothesis tests of a true effect

Imagine an effect whose value in the population is very large, e.g., Cohen's *d* = 1.0. We can label this effect ‘true’ insofar as its value in the population is non-zero. The treatment group has a mean of 1 and the control group a mean of 0, both have an SD of 1, and the data follows a normal distribution. That is, an effect size of Cohen’s *d* = 1.0, given . This population distribution is plotted below in Figure 1. Clear differences between the groups are visible in the form non-overlap of the two distribution curves.

Of course, real world research studies (typically) do not collect data from the whole population but instead draw samples of participants from them. Inferential statistics are then used to make inferences about the population parameter values, aka the existence and magnitude of the true effect. Due to sampling variance, statistical power, etc., the presence of a true effect (i.e., a non-zero effect via a significance test) will only be detected in a proportion of cases. In order to illustrate this fact, we can simulate three experiments drawing samples of participants from this population using a pseudorandom number generator. In each experiment, we sample 13 participants in each of the treatment and control groups, each drawn from these populations. In each experiment, a *t*-test is applied to the simulated data, whose *p*-value is reported in the plot (see Figure 2). The population effect remains to be Cohen’s *d* = 1.0. However, because of statistical power and sampling variance, only some of these experiments detect a the presence of a significant effect (i.e., *p* < .05). Experiments where this non-zero population effect is not detected by the statistical inference test therefore represent False Negatives.

### Null hypothesis tests of a null effect

We can also imagine an effect whose value in the population is zero, i.e., Cohen’s *d* = 0. We can label this a ‘null’ effect. In this case, the population effect for both treatment and the control groups have means of 0. This population distribution is plotted in Figure 3. No differences are noticeable between the groups as one distribution curve completely overlaps and hides the other.

Again, our real world research studies do not collect data from the whole population but rather draw samples of participants from them. Despite the population effect being null, a significant effect will still be detected in some samples due to sampling variance. Cases where a null population effect incorrectly return a significant result on a statistical inference test therefore represent False Positives.

### Long-run False Positive Rates

Following the Neyman-Pearson approach to Null Hypothesis Significance Testing (NHST), statistical inferences about population effects from a given study are informed by our estimates of long-run error rates, as defined by the design, sampling, and analytic choices made by those who ran the study. These values inform how we should interpret the results of a given study in order to estimate the status of the true population effect (e.g., as being true vs. null).

These simulations are specifically concerned with the False Positive Rate (FPR) from frequentist Null Hypothesis Significance Tests, which is heavily influenced by the alpha value used for a given test. Alpha is typically set to 0.05 to establish a long-run probability that, assuming the null hypothesis is true, data at least extreme as those observed would be observed in less than in 5% of cases. This is the definition of a *p* value. *p* values smaller than this cutoff (typically < .05) are frequently referred to as representing ‘statistically significant’ evidence for an effect (although see Lakens REF).

Long run probabilities can be simulated using exactly the same method used in the above examples. Samples are drawn from specified probability distributions using specific parameters, statistical inference tests are applied to the data, and the proportion of cases where the test produces an incorrect inference is then calculated. Whereas the above examples simulated 3 experiments each, in order to produce reliable results, we can simulate a far larger number of experiments (i.e., 1000). The proportion of cases in which a significant result is obtained despite a population effect of 0 equals the False Positive Rate. With an alpha value of 0.05, the False Positive rate should equal the alpha value, assuming that the assumptions of the t-test were met (i.e., random sampling, independence, normality of data in each condition, and homogeneity of variances). Due to the contrived nature of simulations, these assumptions are known to have been met. Results from the simulation confirm that the False Positive Rate is equal to 0.05, as implied by the alpha value (see Appendix 1).

Unfortunately, there are many reasons why the long run the False Positive Rate can be considerably higher than the alpha value. These include simply running multiple significance tests and considering the False Positive Rate of the entire set. That is, increasing the “family-wise” error rate that at least one of the tests will return a False Positive. This has been well-studied in the context of 2-way ANOVA (e.g., REFs from Cramer). Imagine a study comparing the efficacy of an intervention between men and women whose data is analyzed using a 2 (condition: treatment vs control) X 2 (gender: men vs women) ANOVA. This ANOVA produces three *p* values: one testing the main effect for condition, one testing the main effect for gender, and the interaction effect between condition and gender. Assuming that the researchers are willing to make inferences based on any of these three *p* values, the False Positive Rate for the analysis as a whole is now , where *k* is the number of tests run, and assuming the design is fully balanced. Using alpha = 0.05 and *k* = 3, FPR = 0.14. Despite using the common alpha value of 0.05, assuming the null hypothesis is true, nearly three times this proportion of studies (14%) will return a false positive result.

Whereas the False Positive Rate for fully balanced between subjects ANOVA such as this can be

### Estimating the False Positive Rate in multi-way Repeated Measure ANOVAs

These simulations are concerned with ways in which the family-wise false-positive rate in IRAP research often rises above the $a$ value in a way that is not always known to the researcher and/or not made explicit to the reader. As such, the false-positive rate may be greatly inflated in the IRAP literature, and therefore the severity of any given test result (i.e., its diagnosticity of a non-null population parameter) is reduced.

## Simulating the False Positive Rate in the analysis of IRAP data

These simulations are not meant to be exhaustive. The point here is to illustrate the extremely deleterious impact of just a very small subset of a common analytic strategy often employed in IRAP research on the false-positive rate relative to the $a$ value that was employed.

Importantly, these simulations do not attempt to model any (other) form of p-hacking. For example, they do not model conducting a larger set of analyses and selectively reporting results based on significance (e.g., whether or not block order was included in the ANOVA), or changing data processing methods (e.g., trying multiple trial-type inversions, exclusion criteria, trial-type collapsing methods), or optional-stopping in data collection (e.g., checking the significance of results and collecting additional participants until significance is reached). They simply calculate the number of simulated studies that report a significant result, following a single ANOVA and (in some simulations) follow-up t-tests. Any such form of p-hacking (whether done intentionally or unintentionally) would inflate false positive rates even further. See Stefan & Schönbrodt (2022) "Big Little Lies: A Compendium and Simulation of p-Hacking Strategies" for discussion of these strategies and simulations of their impact on the false-positive rate.

### Simulation 1: the Impact of hidden multiplicity in ANOVA on the False Positive Rate.

Cramer et al. (2016) note that it is a common misconception that ANOVA controls for family-wise error rates. Many are still taught that we use ANOVA over, say, multiple t-tests for this reason. The authors demonstrate via simulation study that the false positive rate for a 2X2 between participants ANOVA with alpha = 0.05 is in fact 14%.

In the case of IRAP studies, a modal analysis strategy is to compare performance between two groups on the four IRAP trial-types using a 4(trial type)X2(group) mixed within-between ANOVA. I therefore simulate the false positive rate for such an ANOVA here. Each simulated study uses the modal N used within the IRAP literature (N = 36).

### Simulation 2: the impact of hidden multiplicity in ANOVA, the use of difference from zero t-tests, and between groups t-tests

In addition to such an ANOVA, IRAP studies typically also (1) explore differences between the groups on each trial-type using between-groups t-tests and (2) examine the presence of IRAP effects on each trial type for each group using difference-from-zero t-test. Of course, there is heterogeneity in the (reported) running of tests, and whether running of t-tests was contingent on the results of the ANOVA etc, but this represents a common strategy.

### Impact of hidden multiplicity in ANOVA + difference from zero t-tests + between groups t-tests + generic pattern

Given the results of Hussey & Drake (2020) "The IRAP is not very sensitive to the attitudes and learning histories it seeks to assess", I have reason to believe that there is a generic pattern among IRAP trial-types that is unrelated to the target stimuli used in the task. That is, regardless of what target stimuli are employed, a similar pattern of effect is found - including when the target stimuli are non-words such as CUG and VEC. This generic pattern likely gives rise to incorrect conclusions about the stimulus class rather than merely the generic pattern. E.g., that evidence for "life-positive" implicit attitudes was found rather than merely "the generic pattern of IRAP effects" was demonstrated, unrelated to the target stimuli (life/death).

As such, it is informative to simulate the impact of this generic pattern on the positive rate. In this case, at the level of detecting statistical effects, this rate is no longer a "false" positive rate but a generic positive rate, as it includes both the detection of true-positive and true-null effects. However, for many researchers it still represents a false positive rate if false conclusions (i.e., related to the target stimulus classes) are made on the basis of these results.