Revisiting Factorial Designs: Are Factorial Interaction Tests Really Necessary?

[1]Robert S. Rodger and [2*]Andrew R. Delamater

[1]Dalhousie University, Halifax, Nova Scotia
[2]Brooklyn College, City University of New York

*Contact Author: Andrew R. Delamater
andrewd@brooklyn.cuny.edu

Keywords: Interaction, Factorial Analysis, Post-hoc Contrasts

Abstract

An experimental procedure often used in neuroscience research (and in related disciplines) is the factorial design. Nieuwenhuis et al. (2011, *Nature Neuroscience*, 14(9), 1105-1107) surveyed the literature and observed that researchers frequently analyze their data by performing 'simple effects' tests for each of two (or more) sets of conditions without also performing a factorial interaction test. They argued that only when the factorial interaction test achieves statistical significance can the researcher legitimately claim that an interaction exists in the data, even when separate 'simple effects' tests reveals a difference in one case but not in others. We discuss several ways in which 'simple interactions,' as opposed to 'factorial interactions,' exist and can be evaluated, and conclude that statistical decisions about population parameters (e.g., the $\mu_{ij's}$ in an experiment) should be made only for linearly independent contrasts through them (e.g., $\mu_{11}-\mu_{12} = 0$; $\mu_{11}+\mu_{12}-2\mu_{22} > 0$). Failure to use linearly independent contrasts can lead to *contradictions* or *repetitions* in the conclusions one reaches about population $\mu_{ij's}$ from their statistical hypotheses, and, therefore, should be avoided. We argue here that the factorial form of analysis should not be obligatory when analyzing data from factorial designs. More generally, methods that use a set of linearly independent (preferably mutually orthogonal) contrasts chosen *post-hoc* to evaluate data (e.g., Rodger, 1974; 1975a; 1975b) provide the researcher with the most flexibility in detecting true simple interactive effects, avoid the contradiction and repetition problems, and are, therefore, recommended.

Nieuwenhuis, Forstmann, and Wagenmakers(2011) criticized the form of statistical analysis commonly employed in a variety of types of neuroscience research (and other areas). They observed that investigators often perform some t tests on data (e.g., 'simple effects' tests) without also assessing an 'interaction' test, and asserted that this is bad statistical practice. We agree with Nieuwenhuis et al. (2011) that researchers often do not use their statistical tools appropriately; however, we disagree with their specific set of recommendations and point out alternative ways of conceptualizing what amounts to being good statistical practice in detecting true 'interactive' effects in data.

The crux of our argument is that by conforming to the standard factorial analysis (i.e., performing main effects and interaction tests with follow up simple main effects tests, etc), the researcher will often perform statistical tests that are not optimal for detecting true effects given various obtained patterns of results. This could lead to an unnecessary increase in Type II error rate, and an even more serious problem arises when interaction tests are followed by non-orthogonal simple effects tests. As detailed more below, this practice can lead the investigator to reach contradictory conclusions about their data because of the non-orthogonality of the tests performed. One simple example of this is where an investigator compares three separate experimental conditions (A, B, C) and finds statistically significant differences among them. However, when testing the three individual paired comparisons (hypotheses A-C=0, A-B=0, B-C=0) the researcher finds that A > C, but that A=B and B=C. This is a logical contradiction and it stems from the fact that the three hypotheses are linearly dependent upon one another (i.e., the first hypothesis = the second + the third). This state of affairs creates much uncertainty over the interpretation of data. We suggest that to avoid these problems, one should use a post-hoc method that also ensures a maintained Type I error rate, low Type II error rate, and substantial power (with a suitable sample size calculation performed prior to data collection). Rodger (1974, 1975a, 1975b) introduced a set of procedures that accomplishes these aims and we illustrate below how this approach can lead the investigator to reach unambiguous conclusions with a high degree of statistical power that also avoids the problems inherent in a planned approach such as those dictated by the standard factorial form of analysis.

An Illustration of the Problem with Standard Factorial Analyses

As a representative example of the Nieuwenhuis et al. (2011) argument, consider an experiment in which a set of fear-conditioned mice were tested for responding to a conditioned stimulus paired with foot shock (CS+) and a control stimulus not paired with foot shock (CS-). During a test session, the mice received optogenetic suppression (OpS) of amygdala neurons on some test trials with CS+ and CS- but not on others (NoS) with these stimuli. Over those four conditions, the investigator observed increased freezing to the CS+ relative to CS- when amygdala neurons were not optogenetically suppressed ($P < 0.01$), but no differences between these stimuli when they were ($F < 1.0$, see Figure 1). From this analysis, the investigator concludes that optogenetic suppression of amygdala neurons interferes with the expression of conditioned fear. In this fictitious example, the researchers set their Type I error rate = 0.05 and decided that the population means represented by CS+ and CS- responding under control conditions differ from one another (i.e., $\mu_{11}-\mu_{12} > 0$) because $P < 0.01$, but that they do not differ from one another with optogenetic suppression (i.e., $\mu_{21}-\mu_{22}=0$) because $F < 1.0$.

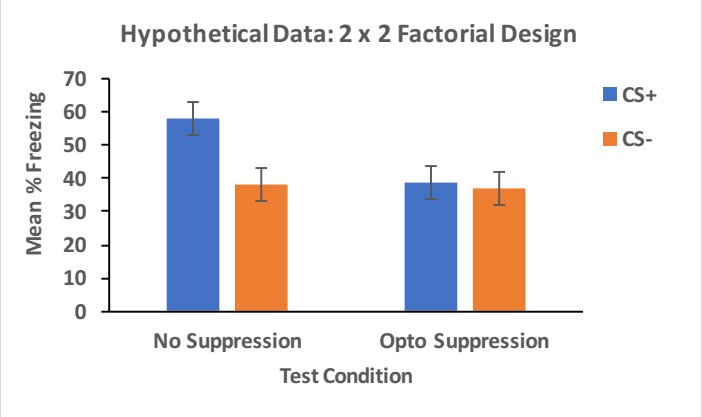

Figure 1: Hypothetical data from a fear conditioning experiment in which a stimulus was paired with foot shock (CS+) or not (CS-) and then tested under conditions in which amygdala neurons were optogenetically suppressed or not.

Nieuwenhuis et al. (2011) argued against this sort of conclusion, without the additional support of an interaction test, specifically examining whether the magnitude of the between-stimulus difference on the control test trials exceeded that displayed when amygdala neurons were suppressed. "To support this claim, they needed to report a statistically significant interaction, … but instead they reported that one effect was statistically significant, whereas the other effect

was not." We assume that Nieuwenhuis et al. (2011) are referring to a 'Factorial Interaction,' but also note that there are other types of 'simple interactions' that may arise in data. For example, when a particular treatment combination gives results that are otherwise atypical, we would argue that this also reflects an interactive effect of those conditions, but the factorial interaction test may not be especially well suited for examining this. We, therefore, disagree with Nieuwenhuis et al. (2011) that a 'Factorial Interaction' test would be required, or even desirable. We elaborate below.

Table 1 shows that the statistical parameters the researchers tested in the example described above can be thought of in the context of a 2x2 factorial design, and Table 2 presents sample means in these four test conditions (also depicted in Figure 1 above):

Table 1: Populations Studied in a 2 x 2 Factorial Design

μ_{ij}	A_1	A_2	$A_.$
B_1	μ_{11}	μ_{12}	$\mu_{1.}$
B_2	μ_{21}	μ_{22}	$\mu_{2.}$
$B_.$	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

Table 2: Possible Sample Means in a 2 x 2 Factorial Design

m_{ij}	CS+	CS−	$m_{i.}$
NoS	58	38	48
OpS	39	37	38
$m_{.j}$ =	48.5	37.5	43

Suppose that in this example, the investigator used a sample size N = 11, and that each mouse was tested IxJ = 2×2 = 4 times with an error mean square = EMS = 275. In this case, the two 'simple effects' t-tests (or equivalent, critical criterion $F_{0.05;1,30} = 4.171 = t^2$) yield:

$$F_1 = N(m_{11}-m_{12})^2/(\nu_1 EMS \Sigma c^2_{ij}) = 11x20^2/(1x275x2) = 8.000 (p=0.008) \{01\}$$

$$F_2 = N(m_{21}-m_{22})^2/(\nu_1 EMS \Sigma c^2_{ij}) = 11x2^2/(1x275x2) = 0.080 (p=0.779) \{02\}$$

From this, the resulting statistical decisions are:

$$X: \mu_{11}-\mu_{12} > 0;\ Y: \mu_{21}-\mu_{22} = 0 \qquad \{03\}$$

The Factorial Interaction from Tables 1 and 2 is equivalent to evaluating the contrast;

$$Z: \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0 \qquad \{04\}$$

giving:

$$F_3 = N(m_{11}-m_{12}-m_{21}+m_{22})^2/(\nu_1 EMS\Sigma c^2_{ij}) = 3.240 \qquad \{05\}$$

which (being less than $F_{0.05;1,30} = 4.171$), supports null hypothesis Z at {04}. Nieuwenhuis et al. (2011) argued that this invalidates conclusion X at {03}.

We disagree with this conclusion. Not only is it not true that X should be invalidated by the acceptance of (or failure to reject) hypothesis Z, but we also point out that the Factorial Interaction null hypothesis (Z) is closely related to hypotheses X and Y, in particular, $Z \equiv X-Y$. In other words, the three contrasts (X, Y, Z) are linearly dependent hypotheses. This can lead to serious problems because any conclusion for the above data that says $Z \leq 0$ is a contradiction, when one accepts that $X > 0$ and $Y = 0$. We maintain that carrying out the Factorial Interaction test in these (X and Y) circumstances is not only unnecessary - it is ill-advised because it is either repetitious (if $Z > 0$), or contradictory (if $Z \leq 0$), with X and Y. This makes the entire statistical claim vacuous.

Other Perspectives

In this example, if the researchers had wished to evaluate another (i.e., their third) contrast, the one that springs to mind is the comparison of the two rows of Tables 1 and 2, i.e.:

$$W: \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} \equiv \mu_{1\cdot} - \mu_{2\cdot} = 0 \qquad \{06\}$$

which is equivalent to the 'rows effect' in a factorial analysis (or, in our example, the main effect of optogenetic stimulation). Note that this hypothesis is not only linearly independent of X and Y above, the three hypotheses are also mutually orthogonal. Performing a set of linearly independent and mutually orthogonal contrasts ensures that the researcher will avoid making statistical decisions that are logically contradictory or repetitious with one another. We believe this to be an essential endpoint for any statistical analysis. In our example, the samples for this contrast give:

$$F_4 = N(m_{11}+m_{12}-m_{21}-m_{22})^2/(\nu_1 EMS\Sigma c^2_{ij}) = 4.000 \quad (p = 0.055) \qquad \{07\}$$

which does not quite reach the $F_{0.05;1,30} = 4.171$ criterion. Under these circumstances we would not advocate changing our α from 0.05 to 0.06, or even suggest that there exists a "marginally significant" main effect. Rather we take this 'small shortage' as an indication that the use of 't-tests for pre-planned contrasts' can sometimes just fall short. Nonetheless, we take the view of Neyman and Pearson (1928a, 1928b) that α is the long-term rate at which 'true' nulls will be rejected (i.e., type-1 errors) and β the long-term rate at which false nulls (with specified non-centrality, say Δ) will be detected. Only by adopting a specific criterion (e.g., $\alpha = 0.05$), and remaining faithful to that criterion, can we expect to erroneously reject true null hypotheses at that rate, and compute a sample size necessary to detect effects of specific sizes at a reasonable rate (e.g., $\beta = 0.95$). We realize, of course, that when we accept a null, it's 'true' value is probably not zero, but it is usually small enough to be considered zero from the point of view of detecting meaningful empirical effects.

The evaluation of the three contrasts X, Y, and W, is not a factorial analysis, but it, nonetheless, shows a Column Effect for Row 1, no Column Effect for Row 2, and no Row Effect; and those three effects are mutually orthogonal.

Another set of mutually orthogonal hypotheses that could be tested, of course, is that which comprises the traditional factorial analysis. In contrast form, these could be written:

Rows: $\mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} = 0$ {08}

Columns: $\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22} = 0$ {09}

Interaction: $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$ {10}

In our example, the Rows and Columns contrasts are the main effects of Optogenetics (Suppression vs No Suppression) and Stimulus (CS+ vs CS-), respectively, while the third is the interaction contrast. These produce the results shown in Table 3.

Table 3: Traditional Factorial Analysis

Source	d.f.	SS	MS	F
Rows	1	1100.0	1100.0	4.000
Cols	1	1331.0	1331.0	4.840
Int	1	891.0	891.0	3.240

Error	30	8250.0	275.0	
Total	33	11572.0		

According to this analysis, and following the recommendations of Nieuwenhuis et al (2011), the only significant effect in these data that could be claimed is the main effect of Stimulus (i.e., Columns) showing that CS+ responding exceeds CS- responding overall. Thus, the obvious pattern of results seen in Figure 1 above is not well captured by this form of analysis, even though the contrasts themselves are mutually orthogonal. However, notice that even had the interaction contrast reached significance, standard statistical practice would then recommend that 'simple effects' contrasts be performed (e.g., Kirk, 1968; Maxwell & Delaney, 1990; Winer, 1962), and at that point the possibility of repetitious or, worse still, contradictory statistical conclusions could be reached because those 'simple effects' tests along with the interaction contrast (e.g., X, Y, and Z), as noted above, are linearly dependent (Z ≡ X-Y).

We take the above as illustrating one of the severe limitations of the planned-contrast and factorial approaches to analyzing data. Instead, we advocate the evaluation of data using a *post-hoc* procedure with the application of appropriate statistical standards. Such an approach, as we will see, offers the researcher much greater flexibility in terms of data analysis, but, importantly, provides them with a statistical tool to more straightforwardly capture the essence of the empirical effect.

Rodger's Method of Post Hoc Evaluation

The reader should not assume that our above demonstration of three different methods of data evaluation indicates that we endorse 'playing around, *post hoc*, with t-tests'. Such a procedure is one that uses a (usually uncalculated) inflated α, with inflation growing as the number of cell means involved increases. One obvious problem with the planned-contrast approach is that researchers might be tempted in view of their results to perform and report the results of 'planned' tests even though those tests were not, in fact, planned. A second problem with the planned approach is that as the number of planned tests performed increases, α can be increasingly inflated. The common practice of correcting against this by adopting a more stringent criterion, e.g., α/(# of tests performed), in order to preserve an experiment-wise error rate

faces other difficulties, the main problem of which is a severe loss in the power to detect true effects.

Rodger (1974; 1975a; 1975b)[1] introduced a method of evaluating data from experiments of the general sort described above that provides the researcher with great flexibility to assess contrasts of interest in view of the pattern of results actually obtained. Because this approach is rather under-appreciated in the literature, we illustrate our arguments by considering this form of analysis. Rather than being confined to examining data with the factorial contrasts noted above, of Rows, Columns, and the Interaction, the researcher is free to select any set of contrasts *post-hoc* providing that they are linearly independent, and, ideally, mutually orthogonal. In this sense, the approach is quite different from evaluating a set of planned contrasts. The approach is extremely straightforward and it avoids the potential problems of reaching contradictory or repetitious conclusions noted above (as well as usually avoiding 'close to significant' results), while at the same time provides a reasonably clear way of evaluating highly informative contrasts that would very likely not be considered otherwise.

Rodger's Method of *post-hoc* contrast evaluation takes simple contrasts as its basis, hence rather than controlling the type-1 error-rate at, for example, $\alpha = 0.05$ for the overall null hypothesis:

$$\text{Ho: } \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_J \qquad \{11\}$$

he sets the average (i.e. 'expected') type 1 error rate at, for example, $E\alpha = 0.05$ for the set of J-1 mutually-orthogonal contrasts that are (*post hoc*) chosen to separate the J values of μ_j into sub-sets. This means that over the long haul the investigator can expect to incorrectly reject 5% of true null contrasts out of a set of J-1 mutually orthogonal contrasts. In order to ensure this expected type I error rate, $E\alpha$, Rodger (1975a) reported new tables of critical values of F with varying ν_1 and ν_2 where $E\alpha = 0.05$ or $E\alpha = 0.01$. The Rodger criterion for the analysis of the data in Table 2, for instance, is $F[E\alpha];\nu_1,\nu_2 = F[0.05];3,30 = 2.023$.

Like his replacement of α for {11} by $E\alpha$ for the average of type 1 errors for J-1, mutually-orthogonal, contrasts; power is conceptualized in terms of the average (i.e., statistical expectation) rate $E\beta$ (instead of power β) for the detection rate of false mutually orthogonal null contrasts. One can then use the approach to calculate a sample size required in order to

detect various sought-after effect sizes to ensure a reasonably high rate of detection. Further details can be read in Rodger (1974; 1975a; 1975b)[1].

To illustrate the approach, our m_{ij} in Table 2 yield the overall F:

$$F_m = N\Sigma(m_{ij}-m..)^2/(\nu_1 EMS) = 4.027 \equiv \Sigma F_h = 4.027 \quad \{12\}$$

in which F_h is the F score computed for an individual contrast:

$$F_h = N\Sigma_{ij}(c_{hij}m_{ij})^2/(\nu_1 EMS\Sigma_{ij}c^2_{hij}) \quad \{13\}$$

One important feature of the approach highlights its difference from the method of planned contrasts noted above. The ν_1 in the divisor of the F_h formula has $\nu_1 = IJ-1$ (= 2 x 2 - 1 = 3 in our example), whereas, for planned contrasts $\nu_1 = 1$. Another important aspect of the approach is that the overall F_m is thought of as being comprised of ν_1 mutually orthogonal F_h parts, ΣF_h. In other words, the sum of the variation within each of the contrasts reflects the total sum of the overall variation in our dataset. This being the case, it follows that the number of rejectable null contrasts, r, that could be found in a contrast set can be easily calculated. In our example, the obtained F_m is large enough to reject:

$$r = [F_m/F[0.05];3,30] = [4.027/2.023] = [1.99] = 1 \quad \{14\}$$

null contrast from a set of mutually orthogonal contrasts. Note that r is not allowed to exceed ν_1 (3 in this case). One straightforward set of mutually orthogonal contrasts chosen for the m_{ij} in our example is:

Table 4: Rodger Analysis Contrasts and Their F_h

	m_{11}	m_{12}	m_{21}	m_{22}			
m_{ij}=	58	38	39	37	Σc^2_{ij}	$\Sigma c_{ij}m_{ij}$	F_h
c_{1ij} =	0	0	1	-1	2	4	$11\times2^2/(3\times275\times2)=0.027$
c_{2ij} =	0	2	-1	-1	6	0	$11\times0^2/(3\times275\times6)=0.000$
c_{3ij} =	3	-1	-1	-1	12	60	$11\times60^2/(3\times275\times12)=4.000$

The contrast set we have chosen here is one of any number of orthogonal sets that could have been chosen *post-hoc* in view of the actual data. The first contrast examines Hypothesis Y above (that with optogenetic suppression CS+ and CS- do not differ), and the second contrast tests whether m_{12} differs from

m_{21} and m_{22} combined (i.e., that CS- responding without optogenetic suppression equals CS+ and CS- responding with optogenetic suppression). Because the F_h scores for these two contrasts do not exceed the criterion (indeed, they are rather close to 0), their null hypotheses are retained and this supports the view that the three low means in our example (38, 39, 37) all come from the same underlying population. The third contrast assesses the 'simple interaction' hypothesis that the atypical m_{11} (CS+ responding without optogenetic suppression) is greater than the remaining three test conditions, and indicates that μ_{11} is greater than the μ_{ij} of the other three test conditions, confirmed by its $F_3 = 4.000 > F[0.05];3,30 = 2.023$ rejecting the null. This particular contrast set reflects well the source of the variation among individual means to justify the conclusion that $\mu_{11} > \mu_{12} = \mu_{21} = \mu_{22}$, a pattern that is quite obvious in the empirical data (see Figure 1).

Because the investigator is free to choose, *post-hoc*, any set of orthogonal contrasts from numerous sets possible, they are better positioned to explore contrasts that make sense in view of the actual data, and, as a result, optimize their chances at detecting the true sources of variation that exist in the dataset. That fact is made especially clear in the present example where one sample mean exceeds three others that differ very minimally. The factorial interaction contrast in our example, we suggest, does not make much sense given this pattern of results. It amounts to comparing the average response given to CS+ without optogenetic stimulation combined with CS- with optogenetic stimulation versus the average response to CS- without optogenetic stimulation combined with CS+ with optogenetic stimulation (see {4} and {10} above). Given the pattern of results depicted above, it is no surprise that the "factorial interaction" contrast does not reach statistical significance. On the other hand, the more straightforward contrast comparing the one large mean against the other three small means does reach significance and clearly illustrates, more directly, how differences between stimuli (in this example) interact with test condition.

Moreover, by adopting Rodger's *post-hoc* approach the researcher can reduce the problem Nieuwenhuis et al. (2011) identify when null contrasts have associated p values that narrowly miss the test criterion. This follows from the fact that researchers can choose contrast sets where the contrast F_h values for those null contrasts that are rejected use up much of the overall $F_{m's}$ worth of variation among the sample means, leaving rather little variation left for nulls that are retained

(as illustrated here). Thus, the researcher would rarely find themselves in a position of retaining a null whose associated F_h value was close to the critical $F[E\alpha];\nu_1, \nu_2$.

Conclusions

We have chosen to illustrate our position with regards to a 2x2 repeated measures factorial experimental design, rather than other popular ones. Space limitations preclude us from elaborating on this point. However, the general logic and conclusions we advance apply equally well to IxJ (or IxJxK, etc) independent groups, randomized blocks, mixed "split plot" factorial designs, as well as to various non-parametric procedures (e.g., Rodger, 1969; in which his critical F values were slightly different from those currently preferred in his 1975a table).

In some sense, we agree with Nieuwenhuis et al. (2011) that statistical treatments of data in neuroscience research (and in many other domains) are not always well-informed by statistical theory. However, we disagree that the proper approach should be one that limits the researcher to test for Factorial Interactions. Indeed, we think it is highly misguided to perform 'simple effects' tests combined with Factorial Interaction tests because such a set of statistical hypotheses are linearly dependent and, consequently, will lead to either contradictions or repetitions in their statistical conclusions regarding population μ_s, and, therefore, they could lead to highly ambiguous statistical conclusions, i.e., by both endorsing and denying the same statement about population parameters. It is not uncommon, for instance, for a researcher to find a statistically significant interaction from such an analysis, only to then find no significant simple main effects of interest. Or, as we illustrate above, individual 'simple effects' tests reveal differences, but the interaction test does not. These situations raise concerns to most researchers. Unfortunately, when these situations arise, investigators all too often adopt what is perceived to be the 'conservative' attitude that the conclusion of interest is not warranted by the data. We suggest that, in these cases, the conclusion of interest may not be warranted by the particular form of data analysis that was, perhaps, rather unwisely chosen in the first place. The very notion that the Factorial form of analysis is more "conservative," and, therefore, preferable, we suggest, has little rationale in sound statistical theory, especially, when other more straightforward approaches exist.

As is apparent, we suggest that just because an experimental design takes on a Factorial form this should not obligate the researcher to adopt the Factorial form of analysis, as suggested by Nieuwenhuis, et al (2011). We are not keen on Factorial Analyses, more generally, because (properly applied) they limit the researcher's freedom to evaluate statistically what may seem to be interesting relationships among the sample means. Many interesting contrasts, in view of the obtained data, would be prohibited by Factorial forms of analysis (e.g., hypothesis c_{3ij} in Table 4 above), and some of those contrasts permitted will frequently lead to "near misses." In order to best capture the overall variation among population means suggested by our data set, such planned approaches will frequently fall short, and, therefore, may prevent the uncovering of true effects. However, appropriate use of *post-hoc* methods, such as Rodger (1974; 1975a, 1975b) will help the researcher avoid these pitfalls and, ultimately, lead the investigator to reach a more cogent set of conclusions about population μ_s that more faithfully reflect the patterns of data actually observed.

One final issue concerns the appropriateness of using a decidedly post-hoc procedure in view of recent concerns over "p-hacking," "HARKing," the replicability crisis, and a general bias towards using planned statistical approaches. We believe that Rodger's method is less susceptible to these types of problems. First, while we agree that treating analyses after the results are known as though they were planned is problematic (Kerr, 1998), because Rodger's approach is a post-hoc procedure it is not susceptible to this particular problem. Second, a further problem (alluded to above) arises when an investigator fails to anticipate the obtained data pattern. In that case, the planned-contrast approach forces the investigator to perform analyses that, in view of the obtained data pattern, may make very little sense. Rodger's post-hoc approach avoids this all too common problem. Third, the approach we advocate has built-in constraints that should protect the investigator from some p-hacking concerns. For instance, the requirement that only ν_1 linearly independent (and, ideally, mutually orthogonal) contrasts be selected when the overall analysis supports $r > 0$ rejectable null contrasts (equation {14} above) places important constraints on the scope of the analysis. The investigator is free to examine any and all contrasts as they wish until they achieve and report a set of mutually orthogonal contrasts which contain r null rejections and $\nu_1 - r$ null acceptances and reflect a sensible scientific interpretation of the data. This should not be confused with "p-hacking" because the investigator is

forced to follow the above set of rules and by doing so should not expect to reject true nulls at a rate higher than $E\alpha$. Thus, we do not believe that there is anything inherently wrong with this post-hoc procedure. Indeed, we think it is preferable over planned-contrast or Factorial Analysis approaches because it enables the researcher to best identify the observed source of true variation among sample means.

Finally, regarding the problem of the failure to replicate research findings, Rodger (1974) noted that statistical power, as standardly defined, severely decreases as the number of means in an analysis, i.e., ν_1, increases. This may sometimes affect a failure to replication where the number of groups has increased. However, because of Rodger's unique way of defining error rate (see above), this same loss of statistical power does not occur with increases in ν_1 when using Rodger's method (see Rodger, 1974). Moreover, Rodger and Roberts (2013) compared the statistical power across different methods of analysis, and found Rodger's method to be among the most powerful in detecting true effects. Thus, smaller sample sizes would be required of this method to detect equivalent true effect sizes compared to other approaches. We agree, though, that it is important for an experimenter to begin their study with a power assessment to determine what sample size would be required to detect a given effect size with a high degree of power.

Acknowledgements

The research reported here was supported by a National Institute on Drug Abuse and the National Institute for General Medical Sciences (SC1 DA034995) grant awarded to ARD. R. S. Rodger is currently an Adjunct Professor with Dalhousie University. The authors gratefully acknowledge Geoffrey Schoenbaum, Matthew Crump, Byron Nelson, Jake Jacobs, and Matthew Gardner for providing helpful comments on earlier versions of this manuscript.

Author Contributions

RSR devised the set of post-hoc statistical methods advocated here in a series of papers from the 1970s. Both RSR and ARD extensively discussed that approach with reference to the Nieuwenhuis et al (2011) paper, and both authors jointly wrote the present paper.

Competing Interests

The authors have no competing interests to declare.

Notes

[1] In addition to Rodger's published articles, a wikiversity site exists describing the main elements of his approach: https://en.wikiversity.org/wiki/Rodger%27s_Method

References

Kerr, N.L. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* **2**, 196-217 (1998).

Kirk, R.E. *Experimental Design: Procedures for the Behavioral Sciences* (Wadsworth Publishing Co., Belmont, CA, 1968).

Maxwell, S.E. & Delaney, H.D. *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (Brooks/Cole Publishing Co., Pacific Grove, CA, 1990).

Neyman, J. & Pearson, E.S. On the use and interpretation of certain test criteria for the purpose of statstical inference: part I. *Biometrika* **20A**, 175-240 (1928a).

Neyman, J. & Pearson, E.S. On the use and interpretation of certain test criteria for the purpose of statistical inference: part II. *Biometrika* **20A**, 263-294 (1928b).

Nieuwenhuis, S., Forstmann, B.U. & Wagenmakers, E-J. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience* **14**,1105-1107 (2011).

Rodger, R.S. Linear hypotheses in 2xa frequency tables. *British Journal of Mathematical and Statistical Psychology* **22**, 29-48 (1969).

Rodger, R.S. Multiple contrasts, factors, error-rate and power. *British Journal of Mathematical and Statistical Psychology* **27**, 179-198 (1974).

Rodger, R.S. The number of non-zero, *post hoc* contrasts from ANOVA and error-rate I. *British Journal of Mathematical and Statistical Psychology* **28**, 71-78 (1975a).

Rodger, R.S. Setting rejection rate for contrasts selected *post hoc* when some nulls are false. *British Journal of Mathematical and Statistical Psychology* **28**, 214-232 (1975b).

Rodger, R. S., Roberts, M. Comparison of power for multiple comparison procedures. *Journal of Methods and measurement in the Social Sciences* **4(1),** 20-47 (2013).

Winer, B.J. *Statistical Principles in Experimental Design* (McGraw-Hill Book Co., New York, NY, 1962).