Richter, M. (in press). A critical comment on residual tests in the analysis of planned contrasts. *Psychological Methods*.

Residual tests in the analysis of planned contrasts: Problems and solutions

Michael Richter

University of Geneva

Author Note

Michael Richter, Department of Psychology, University of Geneva.

I am grateful to Kerstin Brinkmann, Mathieu Chatelain, Guido H. E. Gendolla, Nicolas Silvestrini, Rex A. Wright, and Athina Zafeiriou for comments on an early version of this article.

Correspondence should be addressed to Michael Richter, Geneva Motivation Lab, Department of Psychology, FPSE, University of Geneva, CH-1205 Geneva, Switzerland. E-mail: michael.richter@unige.ch

Abstract

It is current practice that researchers testing specific, theory-driven predictions do not only use a planned contrast to model and test their hypotheses but also test the residual variance (the C+R approach). This analysis strategy relies on work by Abelson and Prentice (1997) who suggested that the result of a planned contrast needs to be interpreted in the light of the variance that is left after the variance explained by the contrast has been subtracted from the variance explained by the factors of the statistical model. Unfortunately, the C+R approach leads to six fundamental problems. In particular, the C+R approach (1) relies on the interpretation of a non-significant result as evidence for no effect, (2) neglects the impact of sample size, (3) creates problems for a priori power analyses, (4) may lead to significant effects that lack a meaningful interpretation, (5) may give rise to misinterpretations, and (6) is inconsistent with the interpretation of other statistical analyses. Given these flaws, researchers should refrain from testing the residual variance when conducting planned contrasts. Single contrasts, Bayes factors, and likelihood ratios provide reasonable alternatives that are less problematic.

Residual tests in the analysis of planned contrasts: Problems and solutions

Planned contrasts enable researchers to conduct tailored statistical tests of their hypotheses (e.g., Abelson, 1995; Hays, 1988; Rosenthal & Rosnow, 1985; Winer, Brown, & Michels, 1991). In contrast to the default ANOVA or GLM tests offered by many statistical software packages as standard analysis tool, planned contrasts have often a higher statistical power (e.g., Myers & Well, 1995) and allow researchers to address their hypothesis with a single test. Proponents of contrast analysis agree on the utility of the method. They disagree, however, regarding the specific analysis strategy. Some authors suggested that only a single planned contrast is needed to demonstrate whether the data support the hypothesis or not (*single contrast approach*; Furr, 2004; Furr & Rosenthal, 2003; Rosenthal & Rosnow, 1985). Other authors argued that a single contrast is not conclusive and that an additional, non-significant test of the residual variance is needed before one may conclude that the data convincingly support the hypothesis (*C+R approach*; Abelson & Prentice, 1997; Brauer & McClelland, 2005; Niedenthal, Brauer, Robin, & Innes-Ker, 2002). The disagreement on the best contrast analysis strategy is also reflected in the empirical literature. Six percent of all empirical papers that were published in 2013 in the *European Journal of Social Psychology*, the *Journal of Experimental Psychology: General*, *Motivation and Emotion*, the *Personality and Social Psychology Bulletin*, and *Psychological Science* used planned contrasts to model and test the predictions. Half of these papers followed the single contrast approach and conducted only the contrast test. The other half adopted the C+R approach and tested additionally the residual sum of squares.

The decision between the two approaches is not trivial given that it has important implications for data interpretation. Imagine that a researcher, Dr. Mustermann, has developed a theory about the time that students spend on preparing an exam. Her theory claims that the time students invest is a function of the expected difficulty of the exam: The higher the expected difficulty, the higher the number of hours that students spend on preparing the exam. To test her theory, Dr. Mustermann runs a study in which she assesses the number of hours that students spend on preparing an easy, a moderately difficult, a difficult, and a very difficult exam. She then models her hypothesis with a planned contrast and conducts the corresponding statistical test. The test is significant. According to the single contrast approach, she would have convincingly supported her prediction. However, according to the C+R approach, she would have to conduct an additional test. She would have to subtract the sum of squares associated with her contrast from the total between-group sum of squares and test the resulting residual sum of squares. Only if this residual sum of squares–the variation explained by exam difficulty that is not captured by her contrast–does not attain statistical significance, she could claim that her data unambiguously support her hypothesis. In this paper, I will compare both contrast analysis strategies, discuss problems associated with the C+R approach, and present alternatives.

Contrast analysis

Contrasts enable researchers to answer their specific research questions with tailored statistical tests. The hypothesis of interest is first translated into contrast weights (see Furr, 2004, for a tutorial on how to translate a research hypothesis into adequate contrast weights). Dr. Mustermann's research question, for instance, could be modeled using a linear contrast with the weights -3 (easy exam), -1 (moderately difficult exam), +1 (difficult exam), and +3 (very difficult exam). The only constraint that the contrast weights have to meet is that they sum up to zero. After having assigned the contrast weights, the sum of squares associated with the contrast is calculated using the following formulas:

$$L = \sum_{i=1}^{k} c_i M_i \tag{1}$$

$$w = \sum_{i=1}^{k} \frac{c_i^2}{n_i} \tag{2}$$

$$SS_{contrast} = \frac{L^2}{w} \tag{3}$$

where k is the number of groups or treatment levels, c_i is the contrast weight of group i, M_i is the mean of group i, and n_i is the number of participants in group i.[1] Using the

[1] Rosenthal and Rosnow (1985) suggested that the harmonic mean of the number of participants in the groups should be used instead of the individual n_i if the n_i are not equal.

data of Dr. Mustermann's first study presented in Table 1, one obtains for Dr. Mustermann's contrast $L = 10.30$, $w = 1.67$, and $SS_{contrast} = 63.65$.

A F test for the contrast can be computed by dividing $MS_{contrast}$ by the associated mean square of error (MS_{error}). Given that all contrasts are associated with a single degree of freedom, $SS_{contrast}$ equals $MS_{contrast}$ and the numerator degrees of freedom of $F_{contrast}$ are always one. The denominator degrees of freedom are the degrees of freedom associated with the MS_{error}. The appropriate MS_{error} is the MS_{error} that is used to test the general effect of the factors involved in the contrast (see Rosenthal & Rosnow, 1985, for a detailed discussion on how to find the appropriate error term). In Dr. Mustermann's case, the associated MS_{error} is the MS_{error} that is used to test the exam difficulty effect in a one-way ANOVA. If the tested prediction is directional, the F test can be replaced by the equivalent t test:

$$F_{contrast} = t^2_{contrast} = \frac{MS_{contrast}}{MS_{error}} \tag{4}$$

Using the data of Dr. Mustermann's first study, one obtains $F_{contrast}(1,\, 44) = 121.77$ (or $t_{contrast}(44) = 11.04$), $p < .001$, $MS_{error} = 0.52$.

The C+R approach

The variance explained by any single factor of a fixed-effects analysis of variance (ANOVA) can be decomposed into a set of k-1 orthogonal contrasts (k corresponds to the number of factor levels or groups). Each contrast of the set is associated with one degree of freedom of the ANOVA factor and explains a unique part of the variance explained by the factor. The k-1 orthogonal contrasts explain together all variance explained by the factor. In other words, the total variation between the different levels of a factor–the sum of squares associated with the factor–can be partitioned into the specific patterns of variation predicted by the orthogonal contrasts (see Equation 5). Each contrast predicts a certain ranking of factor levels and explains a part of the total sum of squares.

$$SS_{factor} = SS_{contrast\,1} + \cdots + SS_{contrast\,k-1} \tag{5}$$

Dr. Mustermann's manipulation of exam difficulty can be conceptualized in a one-way ANOVA as a between-persons factor with three degrees of freedom. One possibility to decompose this factor into orthogonal contrasts is to use a set consisting of a linear contrast (contrast weights -3, -1, +1, and +3)–corresponding to Dr. Mustermann's contrast of interest–a quadratic contrast (contrast weights +1, -1, -1, and +1), and a cubic contrast (contrast weights -1, +3, -3, +1). However, this set constitutes only one of many possible sets of orthogonal contrasts that explain the variance associated with the exam difficulty factor. For any first contrast, orthogonal contrasts can be found. For instance, a contrast with the weights +1, +1, -1, and -1, a contrast with the weights -1, +1, -1, and +1, and a contrast with the weights -1, +1, +1, and -1 would also constitute a set of orthogonal contrasts that accounts for all of the variance explained by the exam difficulty factor.

The contrast that a researcher uses to model and to test her or his prediction thus explains only a part of the total variation between the groups. The remaining variance is explained by the other contrasts of the set of orthogonal contrasts. Abelson and Prentice (1997) used the terms ”residual” and ”residual variance” to refer to this amount of variance suggesting that the total explained variance should be partitioned into variance explained by the researcher's contrast and residual variance as shown in Equation 6. Abelson and Prentice (1997) thus used the term ”residual” to refer to explainable, non-error variance and not, like in many other statistical approaches, to refer to error variance.

$$SS_{factor} = SS_{contrast} + SS_{residual} \tag{6}$$

Abelson and Prentice claimed that an analysis of the residual variance is crucial for a valid interpretation of the outcome of the contrast test and that any contrast test should be followed by a test of the residual variance. They argued that one may miss systematic patterns of interest if one only tests the contrast. According to Abelson and Prentice, a significant contrast and a non-significant residual–they call this pattern a canonical outcome–show that the data fit the predictions and that deviations from the predicted pattern are random. If both tests are significant–an ecumenical outcome in

Abelson and Prentice's terms–the data pattern resembles the theoretical predictions but there is additional systematic variation beyond the expected pattern. Abelson and Prentice thus differentiated between a pattern of results that provides full, parsimonious support for the predictions and a pattern that provides less support. Niedenthal (Niedenthal et al., 2002) and Brauer (Brauer & McClelland, 2005) took one step further by suggesting that only a significant contrast that is accompanied by a non-significant residual enables researchers to conclude that their hypothesis has been confirmed. If the residual is significant, the researcher failed to provide evidence for her or his hypothesis.

The calculation of a test of the residual is straightforward. First, $SS_{residual}$ is calculated by subtracting $SS_{contrast}$ from SS_{factor}. $SS_{contrast}$ has already been computed for the contrast test. SS_{factor} may be calculated by hand using the formulas provided in various textbooks (e.g., Hays, 1988; Winer et al., 1991) but it might be easier to calculated an ANOVA using one of the available statistics software packages and to copy the sum of squares from the output. $MS_{residual}$ is then calculated by dividing $SS_{residual}$ by the associated *k-2* degrees of freedom. Finally, an F test is computed using

$$F_{residual} = \frac{MS_{residual}}{MS_{error}} \tag{7}$$

For Dr. Mustermann's study, one obtains $SS_{contrast}$ = 63.65, SS_{factor} = 74.49, $SS_{residual}$ = 10.84, and $MS_{residual}$ = 5.42. The F test results in $F_{residual}(2, 44) = 10.36$, $p < .001$. Dr. Mustermann thus obtained significant results for the contrast and the residual. Depending on the contrast analysis strategy that she favors, she would either conclude that the data support her prediction (single contrast approach) or conclude that the data do not fully support the prediction (C+R approach).

Problems associated with the C+R approach

Analyzing the residual sum of squares as suggested by Abelson and Prentice might seem reasonable. Unfortunately, the C+R approach leads to six major problems. Two of the problems–the first two that I will present–are not specific to the C+R approach. They reflect general shortcomings of *p* value based hypothesis testing and are common to all *p* value based procedures. The other four problems are consequences of the logic of the C+R approach.

Problem 1: The C+R approach asks researchers to interpret non-significant results as evidence for no effect

The C+R approach requires researchers to demonstrate with a non-significant residual test that there is no systematic variance beyond the pattern modeled by their contrast. Researchers are thus asked to interpret a non-significant statistical test as evidence for no effect. A large number of authors questioned this interpretation of non-significant tests arguing that failing to reject the null hypothesis does not imply that one has found evidence in favor of it (e.g., Aberson, 2002; Dixon, 2003; Johansson, 2011; Nickerson, 2000; Wagenmakers, 2007). I will not reiterate all the concerns that have been raised but I will briefly elaborate on two points that demonstrate why interpreting a non-significant test as evidence for no effect is disputable.

First, *p* values are uniformly distributed if there is no true effect (e.g., Rouder, Morey, Speckman, & Province, 2012, 2009). All *p* values are equally likely if the null hypothesis of no effect is true. For instance, it is as likely to obtain a high *p* value in the range from .95 to 1 as to obtain a low and significant *p* value (in the range from 0 to .05). Consequently, a high and non-significant *p* value provides as much evidence for the hypothesis that there is no systematic residual variance as a low and significant *p* value. Showing that the test of the residual yields a particular, non-significant *p* value (e.g., .90) does not provide more evidence in favor of no systematic residual variance than showing a significant *p* value (e.g., .03).

Second, small and non-significant *p* values are, under certain conditions, more likely if there is a small true effect than if the null hypothesis of no effect is true (e.g., Hung, O'Neill, Bauer, & Köhne, 1997; Rouder et al., 2012; Sellke, Bayarri, & Berger, 2001). If there is a true non-zero effect, the *p* value distribution is right-skewed and its shape depends on sample size and the size of the true effect. The higher the sample size and the higher the size of the true effect, the more likely are low *p* values compared to high *p* values. A consequence of this characteristic of the *p* value distribution under a non-zero effect is that the probability of finding a small but non-significant *p* value varies with sample size and the size of the true effect. In contrast, the probability of finding a small but non-significant *p* value is independent of sample size if the null hypothesis of no effect is true. For instance, the probability of obtaining a *p* value between .05 and .10 is always 5% if there is no true effect. If there is a true non-zero effect, the probability of finding a *p* value within this range will be smaller than 5% if the true effect is large or if sample size is large. However, if the true effect is small or if sample size is low, the probability of finding a *p* value between .05 and .10 will be higher than 5% (Sellke et al., 2001). Small, non-significant *p* values can thus be more likely if there is a non-zero true effect than if there is no true effect. Consequently, non-significant *p* values may provide more evidence for a true effect (or systematic residual variance) than for no effect (or no systematic residual variance).

It follows that non-significant *p* values do not necessarily constitute evidence for no systematic residual variance–as suggested by the C+R approach. They may even constitute more evidence for a small amount of systematic residual variance than for no systematic residual variance. It is of note that the problematic interpretation of non-significant results as evidence for no effect is not unique to the C+R approach. Many statistical procedures interpret non-significant *p* values as evidence for no effect or no difference. For instance, tests for normality that are conducted to check model assumptions are often interpreted in this manner. A non-significant result is interpreted as evidence that the population distribution is not different from a normal distribution.

Problem 2: The outcome of the residual test depends on sample size

According to the C+R approach, testing the residual provides information about the performance of the contrast. A significant residual is interpreted as evidence that the contrast performed poorly leaving some systematic variance unexplained. A problem for this interpretation arises from the relationship between *p* values, effect sizes, and sample size. For a given (non-zero) effect size, the *p* value is a function of sample size: Small sample sizes lead to high *p* values, whereas large sample sizes result in low *p* values (e.g., Hung et al., 1997). This also holds for the test of the residual variance. If the residual variance is associated with a non-zero population effect, the probability of getting a significant residual test increases with increasing sample size.

The following example illustrates this problem. Imagine that a researcher manipulates a variable across three levels and predicts a linear effect across the three groups. The linear contrast weights would be -1, 0, and +1. The between-groups variance that would not be captured by this linear contrast, that is, the residual variance, would be associated with a second, quadratic contrast (contrast weights +1, -2, and +1) (e.g., Rosenthal & Rosnow, 1985). Let us also assume that the true (population) relationship between the manipulated variable and the dependent variable is a combination of a strong linear effect and a small quadratic effect. In this case, the likelihood of finding a significant linear effect as well as the likelihood of finding a significant quadratic (residual) effect would both increase with increasing sample size. The results and their interpretation would strongly depend on the size of the collected sample. If the sample size was small, the contrast may be significant but the residual variance may fail to reach significance. The researcher would be allowed to conclude that the data convincingly support the predictions. If the sample size was large, both tests probably would be significant and the researcher would not be allowed to conclude that the data provide full support for her or his prediction.

For researchers following the C+R approach, two problems arise from the described relationship among *p* values, effect sizes, and sample size. First, whether a researcher can provide full support for her or his hypothesis depends on sample size. The outcome of the statistical analysis (canonical vs. ecumenical) does not only depend on the performance of the researcher's hypothesis but also on the size of the collected sample. Given that a large sample size will render even small amounts of residual variance significant, it will be almost impossible to provide full support (i.e., a canonical outcome) with a large sample size. Second, the C+R approach confronts researchers with an unsolvable approach-avoidance conflict. On the one hand, they are motivated to collect few data to keep the likelihood of finding a significant residual low. On the other hand, they are interested in conducting a contrast test that has a high statistical power to detect true effects and are thus motivated to have a large sample size.

Problem 3: The meaning of type I and type II error differs between the contrast test and the residual test

A central motivation for hypothesis testing is researchers' interest in avoiding erroneously claiming that they have found support for their predictions. Tests that follow the Neyman-Pearson paradigm satisfy this interest by offering a tool that enables the control of the long-term rate of making a wrong decision. Within this paradigm, researchers formulate two mutually exclusive hypotheses (the null and the alternative hypothesis) and control two types of long-term error rates: the long-run probability of rejecting the null hypothesis when it is true (type I error) and the long-run probability of not rejecting the null hypothesis when it is false (type II error). Given that in most cases the alternative hypothesis corresponds to the researcher's hypothesis of interest, controlling type I error offers researchers a mean to control the long-run probability of erroneously claiming that they found support for their predictions.

The single contrast approach (and the contrast part of the C+R approach) is in line with this logic. Rejecting the null hypothesis if it is true corresponds to erroneously claiming that one has found support for one's prediction. Correspondingly, researchers can control the long-run probability of erroneously claiming that the data support their prediction by setting the type I error rate of the contrast test. However, to control the same type of error in the residual test, one has to control type II error instead of type I error. Researchers following the C+R approach expect and predict that their contrast explains all systematic variance and that there is no residual variance. They thus predict that the null hypothesis of no effect is true when testing the residual. Correspondingly, rejecting the null hypothesis when it is true (type I error) has a different meaning than in the contrast test. It corresponds to erroneously rejecting the researcher's hypothesis and not to erroneously claiming that one has found support for it. To control the long-term rate of erroneously claiming that one has found support for no residual variance, researchers have to control the type II error rate of the residual test instead of controlling type I error rate. Given that type II error control requires researchers to predict a specific, non-zero effect for the alternative hypothesis, this leads to an awkward situation: Researchers expect and predict no residual effect but they nevertheless have to formulate a prediction about a non-zero effect to control type II error rate.

There is also a practical problem. Most funding agencies ask their applicants to run a priori power analyses to determine required sample sizes. As far as I know, there is no commercial software package that offers a tool that controls at the same time the type I error rate of one test and the type II error rate of another test. Researchers thus have to develop their own tool that enables them to conduct the required power analysis and to determine the sample size needed to detect a canonical outcome.

Problem 4: Residual tests often examine effects that lack a meaningful interpretation

Brauer and McClelland (2005) argued that it is important to test the residual variance because one might miss important information if one does not test it. A significant test of the residual is interpreted as evidence that there are other effects that are worth exploring. Given that the residual variance may be associated with effects that lack a meaningful interpretation, it is questionable whether a significant residual test implies in general that it is important to explore the variance associated with the residual. Dr. Mustermann's linear contrast (contrast weights -3, -1, +1, and +3) constitutes together with a quadratic contrast (weights +1, -1, -1, and +1) and a cubic contrast (-1, +3, -3, +1) a set of orthogonal contrasts that accounts for all variance that is explained by the manipulated variable (e.g., exam difficulty). The variance that is not explained by the linear contrast, the residual variance, thus includes variance associated with the quadratic and the cubic effect. It may be possible to interpret the quadratic contrast in a reasonable manner but I cannot think of a meaningful interpretation of the cubic contrast. Why should students invest less time in preparing a difficult exam than in preparing an easy exam but invest more time in preparing a very difficult exam than in preparing an easy exam? Why should they spend more time on preparing a very difficult exam than on preparing a difficult exam but spend less time on preparing a very difficult exam than on preparing a moderately difficult exam? This example demonstrates that the residual variance may be associated with patterns that lack a meaningful interpretation. I doubt that it is fruitful to examine effects that cannot be interpreted in a meaningful manner. Examining effects that cannot be interpreted does not seem to be helpful for advancing the understanding of a phenomenon.

Problem 5: The C+R approach may lead researchers to erroneously conclude that their prediction offers the best explanation of the data

The C+R approach does not only suggest that a significant residual indicates that there are other effects worth exploring, it also suggests that a non-significant residual test means that other (residual) effects are unimportant. There is a risk that researchers misinterpret this notion. They may be inclined to interpret a non-significant residual as evidence that all other effects are negligible and that their prediction provides the best explanation of the data. Given that meaningful alternative hypotheses may share variance with both the contrast and the residual, this conclusion is not warranted.

Imagine that there is an alternative theory to Dr. Mustermann's theory. Drawing on the idea that there is an upper limit of the time that students are willing to invest for an exam, the alternative theory postulates that students disengage and do not invest any time in preparing an exam if exam difficulty is too high. Applying this theory to Dr. Mustermann's manipulation of exam difficulty, one could predict that preparation time increases with exam difficulty across the first three exam difficulty conditions. In the fourth condition where exam difficulty is very high, participants should disengage because of the very high exam difficulty and, correspondingly, preparation time should be very low in this condition. This prediction of the alternative theory may be modeled and tested using the contrast weights -3, +1, +5, and -3 (a tutorial on how to translate theoretical predictions into contrasts can be found in Furr, 2004).

Let us suppose that Dr. Mustermann runs a second study collecting the data presented under Study 2 in Table 1. Analyzing her data, Dr. Mustermann finds a significant linear contrast, $F(1, 44) = 4.20$, $p = .046$, $MSE = 0.41$, and a non-significant residual, $F(2, 44) = 2.66$, $p = .08$. Drawing on the C+R approach, she might now be inclined to state that the statistical analyses show that her theory provides the best explanation of the data. However, a test of the alternative prediction would also be significant, $F(1, 44) = 6.35$, $p = .02$. A comparison of the variance that is explained by the two competing hypotheses reveals that the alternative hypothesis performs even better than Dr. Mustermann's hypothesis ($SS_{alternative} = 2.62$, $SS_{Mustermann} = 1.73$). Dr. Mustermann's canonical outcome–a significant contrast and a non-significant residual–does not imply that her theory provides the best explanation of the relationship between exam difficulty and preparation time. Given that researchers are normally interested in demonstrating the superiority of their theoretical account, I am afraid that the C+R approach may lead researchers to erroneously interpret a canonical outcome as evidence that their explanation is the best one.

Problem 6: Researchers do not (want to) consistently apply the C+R logic

The C+R approach asks researchers to demonstrate a significant contrast and a non-significant residual. Only if they can show this pattern, they are allowed to conclude that the data convincingly support their prediction. Consider the application of this reasoning to other kinds of statistical analyses, a conventional 2 x 2 between-persons ANOVA, for instance. In a 2 x 2 ANOVA the default tests of the two main effects and the interaction effect constitute a set of orthogonal contrasts that accounts for all explained variance (contrast weights are -1, -1, +1, and +1 for the first main effect, -1, +1, -1, and +1 for the second main effect, and -1, +1, +1, and -1 for the interaction). If a researcher predicts an interaction, the C+R logic would require her or him to show a significant interaction without a significant residual. The test of the variance associated with the main effects should thus be non-significant. If the researcher finds a significant interaction but also significant main effects, the outcome would be an ecumenical one. The researcher would not be allowed to conclude that the data provide full support for the predictions. Following Brauer and McClelland (2005), the researcher would even need to conclude that she or he failed to provide evidence for the predicted effect.

I know many studies where researchers stated that significant main effects were qualified by an interaction but I do not know any study where a researcher toned down the interpretation of a predicted interaction because of the presence of significant main effects. According to the C+R approach, researchers should do that. Whenever the significant predicted effect is accompanied by a significant other, orthogonal effect, researchers failed to provide full support for their predictions (or even failed to provide any evidence for the predicted effect according to Brauer & McClelland, 2005). The C+R approach thus suggests an analysis strategy that researchers do not consistently apply to other kinds of statistical analyses. It is questionable whether it is reasonable to adopt a strategy for the analysis of planned contrasts that is not applied to other statistical analyses.

Solutions

As outlined in the preceding sections, the C+R approach is associated with several serious drawbacks. Fortunately, there are alternatives that avoid many of the problems associated with the C+R approach. I will present in the following the single contrast approach, likelihood ratios, and Bayes factors as alternatives to the C+R approach and I will discuss their performance with respect to the six described problems.

The single contrast approach

Instead of using the C+R approach, researchers could follow the single contrast approach advocated by Rosenthal, Rosnow, and others (e.g., Furr, 2004; Furr & Rosenthal, 2003; Rosenthal & Rosnow, 1985). They could test only the contrast that models their hypothesis of interest and refrain from conducting a statistical test of the residual. This approach also corresponds to the approach that many textbooks introduce as standard tool for the comparison of group means (e.g., Hays, 1988; Maxwell & Delaney, 2004; Winer et al., 1991). As explained in the preceding sections, testing the residual does not provide information that researchers testing specific hypotheses are interested in. It does neither enable them to demonstrate that their hypothesis provides the best explanation of the data (problem 5), nor does it enable researchers to evaluate the performance of their hypothesis by showing that the variance that is not explained by the contrast is negligible (problems 1 and 2). Researchers interested in testing a single, theory-driven hypothesis will, consequently, not lack important information if they do not test the residual. Moreover, not testing the residual will allow researchers to get around the problem related to the meaning of type I and type II error (problem 3) and will render problems 4 and 6 obsolete. The single contrast approach also avoids to some extent problem 5. Given that it does not suggest that testing residual variance allows the researcher to demonstrate that the variance that is not captured by her or his prediction is negligible, researchers will be less at risk for concluding that their explanation provides the best explanation of the data.

Even if the single contrast approach avoids many of the problems associated with the C+R approach, both approaches also have some problems in common. The problems described in the preceding sections that reflect general problems of *p* value based hypothesis testing apply to any procedure that relies on *p* values–including the single contrast approach. Adopting the single contrast approach would therefore not avoid problem 1. A researcher testing a hypothesis that predicts no difference between conditions would have to interpret a non-significant result as evidence for no difference. The dependency on sample size (problem 2) and its consequences are also common to the single contrast approach and the C+R approach. Researchers using the single contrast approach to test their hypothesis that variable A has an impact on variable B might be inclined to aim for a large sample size to increase the probability of finding a significant effect. If the sample size is large enough, even the smallest effect will become significant. For a researcher predicting no relationship between the two variables, it might be more tempting to go for a small sample size to increase the probability of getting a non-significant result. In sum, adopting the single contrast approach enables researchers to avoid some of the problems associated with the C+R approach but it does not solve the problems related to the underlying statistical framework. Table 2 provides a summary of the performance of the single contrast approach with respect to the six problems.

Likelihood ratios and Bayes factors

Bayes factors and likelihood ratios are measures of evidence that enable researchers to provide evidence for their hypotheses without leading to the problems associated with the C+R approach. Both measures contrast the probability of the observed data under one hypothesis (or model), $p(D|H_1)$, with the the probability of the results under a second hypothesis, $p(D|H_2)$. Their interpretation is straightforward. A Bayes factor or a likelihood ratio greater than one implies that the observed results are more likely under hypothesis 1 than under hypothesis 2. A value smaller than one implies that the results are more likely under hypothesis 2 than under hypothesis 1. For instance, a Bayes factor or a likelihood ratio of two indicates that the observed results are twice as likely under hypothesis 1 as under hypothesis 2. A Bayes factor or a likelihood ratio of exactly one indicates that the observed results are as likely under the two hypotheses and that the evidence does not support either hypothesis over the other. To facilitate communication some authors suggested descriptive categories for Bayes factors and likelihood ratios. For instance, Royall (1997) suggested that likelihood ratios of less than eight should be interpreted as weak evidence, likelihood ratios between eight and 32 as moderate evidence, and likelihood ratios of 32 or more as strong evidence. Raftery (1995) suggested a similar classification for Bayes factors. According to his classification, Bayes factors represent weak evidence if they are between one and three, positive evidence if they are between three and 20, strong evidence if they are between 20 and 150, and very strong evidence if they are 150 or higher. However, it is of note that these descriptive categories are only useful for communication purposes and do not represent qualitative differences in the strength of evidence.

Only two of the described problems (problem 2 and 5) apply to Bayes factors and likelihood ratios. Like *p* values, Bayes factors and likelihood ratios vary with sample size. For instance, Rouder et al. (2009) showed that–if there is a small true effect–the Bayes factor favors the null hypothesis of no effect for small sample sizes but favors the alternative hypothesis for high sample sizes. Researchers aiming at providing evidence for no effect might thus be inclined to keep sample size low, whereas researchers aiming at providing evidence for an effect might be interested in having a large sample. Like the C+R approach, Bayes factors and likelihood ratios are vulnerable to misinterpretations. Researchers might be inclined to erroneously interpret a high Bayes factor or likelihood ratio as evidence for the general superiority of their hypothesis or model (problem 5). However, Bayes factors and likelihood ratios reflect the relative evidence in favor of one hypothesis compared to a second hypothesis. They only enable conclusions regarding the relative performance of the two compared hypotheses. They do not enable conclusions regarding the performance of the compared hypotheses in relation to other hypotheses. The comparison of two hypotheses may result in a high Bayes factor or likelihood ratio–strongly favoring one of the two hypotheses over the other–even if the two hypotheses perform poorly when compared to a third hypothesis.

Bayes factors and likelihood ratios provide, however, a solution to the other four problems. They enable researchers to demonstrate evidence for no effect by comparing the null hypothesis of no effect with any other hypothesis (problem 1). Given that Bayes factors and likelihood ratios are not concerned with long-run error control problem 3 does not apply. They also attenuate problem 4 and 6. Given that researchers need at least two hypotheses to calculate a Bayes factor or a likelihood ratio, they are forced to reflect upon alternative hypotheses or models. Researchers may choose to compare their primary hypothesis with the null hypothesis but it might be more likely that using Bayes factors and likelihood ratios will encourage researchers to compare meaningful hypotheses. Table 2 summarizes how Bayes factors and likelihood ratios perform with respect to the six described problems.

There are excellent papers that introduce Bayes factors and likelihood ratios and that elaborate on how these measures may replace *p* value based hypothesis testing (e.g., Blume, 2002; Glover & Dixon, 2004; Goodman, 1999; Masson, 2011; Wagenmakers, 2007). These papers explicitly discuss how Bayes factors and likelihood ratios can be used to examine hypotheses that are normally tested using the C+R approach or the single contrast approach. In particular the papers by Glover and Dixon (2004) and Masson (2011) provide tutorials that are very accessible so that even researchers who are not familiar with Bayesian or likelihood statistics will encounter no problems in understanding the described methods and in applying them to their own work. Drawing on the work of these authors, I will briefly demonstrate in the following how Bayes factors–to be precise, an approximation of Bayes Factors–and likelihood ratios can be used to address Dr. Mustermann's research questions.

Likelihood ratios and approximate Bayes factors can be obtained by comparing the unexplained variation of one model with the unexplained variation of a second model. Researchers who have conducted a classical contrast analysis have already computed all the information that is needed to calculate the two indices. Likelihood ratios can be obtained with the equation

$$\lambda = \left(\frac{\text{model 1 unexplained variation}}{\text{model 2 unexplained variation}}\right)^{\frac{n}{2}} \tag{8}$$

where n corresponds to sample size (Glover & Dixon, 2004). If Dr. Mustermann were interested in comparing her hypothesis with the null hypothesis, she could compare her linear contrast with a model that predicts no differences among the exam difficulty factor levels. The null hypothesis assumes that exam difficulty has no effect and, consequently, the variation that is not explained by the null model equals the sum of $SS_{\text{exam difficulty factor}}$ and SS_{error}. Dr. Mustermann's linear contrast leaves only $SS_{residual}$ and SS_{error} unexplained. Applying Equation 8 to the data of Study 1, one obtains

$$\begin{aligned}\lambda &= \left(\frac{SS_{\text{exam difficulty factor}} + SS_{\text{error}}}{SS_{\text{residual}} + SS_{\text{error}}}\right)^{\frac{n}{2}} \\ &= \left(\frac{74.49 + 23.00}{10.84 + 23.00}\right)^{\frac{48}{2}} \\ &= 10.68 \times 10^{10}\end{aligned} \tag{9}$$

Given that more complex models (i.e., models with more free parameters) always fit the data better than less complex models, the likelihood ratio should be corrected for differences in model complexity before interpreting it. Hurvich and Tsai (1989) suggested the following equation to correct for model complexity:

$$\lambda_{corrected} = \exp\left[k_1\left(\frac{n}{n-k_1-1}\right) - k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda \tag{10}$$

where k_1 is the number of free parameters of model 1, k_2 is the number of free parameters of model 2, and n is the sample size.[2] Using this correction, one obtains a $\lambda_{corrected}$ of $3.42{\times}10^{10}$. Dr. Mustermann's data are thus $3.42{\times}10^{10}$ times more

[2] The formula proposed by Hurvich and Tsai (1989) is only one of several corrections for model complexity that have been proposed (Glover & Dixon, 2004, for a short discussion).

likely–after correcting for model complexity–given her hypothesis than given the null hypothesis. This provides evidence that Dr. Mustermann's model should be preferred over the null model and could be interpreted as strong evidence in favor of Dr. Mustermann's hypothesis (e.g., Royall, 1997). Comparing Dr. Mustermann's hypothesis with the null hypothesis using the data of her second study results in a $\lambda_{corrected}$ of 2.28. The data of Study 2 also favor Dr. Mustermann's hypothesis over the null hypothesis. However, the evidence is much weaker than the evidence provided by the data of Study 1.

Dr. Mustermann could also compute a likelihood ratio to compare the two components that are important in the C+R approach. She could calculated a likelihood ratio that compares her contrast with a model that explains the residual variance. Such a residual model would include all the contrasts that constitute together with Dr. Mustermann's contrast a set of orthogonal contrasts. Given Dr. Mustermann's linear contrast, the residual model would include the prediction of a quadratic and a cubic contrast. She could thus compute a likelihood ratio to compare her contrast with a model that includes a quadratic and a cubic contrast. However, given that there is probably no meaningful interpretation of a model that predicts that exam difficulty has at the same time a quadratic and a cubic effect on exam preparation time (see the discussion of problem 4), Dr. Mustermann might not be interested in this comparison and refrain from computing a likelihood ratio that compares her contrast with the residual. She might be more interested in comparing her model with another meaningful model. For instance, she could compute a likelihood ratio to compare her model with the alternative theory presented in the discussion of problem 5. Applying Equation 8 to the data of Dr. Mustermann's second study, one obtains

$$\begin{aligned}\lambda &= \left(\frac{\text{alternative contrast unexplained variation}}{\text{Mustermann contrast unexplained variation}}\right)^{\frac{n}{2}} \\ &= \left(\frac{SS_{\text{residual alternative}} + SS_{\text{error}}}{SS_{\text{residual contrast}} + SS_{\text{error}}}\right)^{\frac{n}{2}} \\ &= \left(\frac{1.31 + 18.15}{2.20 + 18.15}\right)^{\frac{48}{2}} \\ &= 0.34\end{aligned} \tag{11}$$

The data that Dr. Mustermann obtained in Study 2 are thus $1/0.34 = 2.94$ times more likely given the alternative model than given Dr. Mustermann's model. Given that both models are of the same complexity, it is not necessary to correct for differences in model complexity. Drawing on the likelihood ratio, Dr. Mustermann would thus conclude that the data provide (weak) evidence in favor of the alternative model and against her theory.

Computing an approximation of the Bayes factor using the Bayesian information criterion (BIC) is as straightforward as computing a likelihood ratio (see Masson, 2011; Wagenmakers, 2007). Again, the variance that is not explained by one model is compared with the variance that is left unexplained by a second model. The approximate Bayes factor is calculated using

$$\triangle\text{BIC} = n \ln\left(\frac{\text{model 1 unexplained variance}}{\text{model 2 unexplained variance}}\right) + (k_1 - k_2)\ln(n) \tag{12}$$

and

$$\text{BF} \approx \text{e}^{(\triangle\text{BIC})/2} \tag{13}$$

where k_1 is the number of free parameters of model 1, k_2 is the number of free parameters of model 2, and n is the sample size (Masson, 2011). The approximate Bayes factor can be directly interpreted as relative evidence for the two models. Alternatively, the approximate Bayes factor can be used to obtain posterior probabilities for the two competing models using

$$p_{\text{BIC}}(\text{H}_1|\text{D}) = 1 - \frac{\text{BF}}{\text{BF}+1} \tag{14}$$

and

$$p_{\text{BIC}}(\text{H}_2|\text{D}) = \frac{\text{BF}}{\text{BF}+1} \tag{15}$$

Applying the preceding equations to compare the evidence for Dr. Mustermann's contrast and the null hypothesis using the data of Study 1, one obtains $BF = 1.55{\times}10^{10}$, $p(\text{null hypothesis}|\text{D}) < .001$, $p(\text{contrast hypothesis}|\text{D}) > .99$. Using the data of Study 2 for the same comparison results in $BF = 1.03$, $p(\text{null hypothesis}|\text{D}) = .49$, $p(\text{contrast hypothesis}|\text{D}) = .51$. Using Raftery's (1995) descriptive categories, one would interpret the result of Study 1 as very strong evidence in favor of Dr. Mustermann's hypothesis relative to the null hypothesis. The results of Study 2 did not provide conclusive evidence in favor of Dr. Mustermann's hypothesis. Using the data of Study 2 to compare Dr. Mustermann's model with the alternative model presented in the discussion of problem 5, one obtains $BF = 0.34$, $p(\text{Mustermann's hypothesis}|\text{D}) = .26$ $p(\text{alternative hypothesis}|\text{D}) = .74$. This could be interpreted as weak evidence in favor of the alternative model.

Conclusion

In the preceding sections, I elaborated on several problems and consequences of the C+R approach and discussed alternatives. I pointed out that exploring the residual variance does not allow researchers to demonstrate that their hypothesis provides the best explanation of the data (problem 5). It also does not enable researchers to evaluate the performance of their hypothesis by showing that the variance that was not captured by the contrast is negligible (problems 1 and 2). Moreover, the C+R approach does not provide a reliable tool to detect other meaningful effects (see problems 2 and 4). Instead of providing benefits, testing the residual comes with several drawbacks. Researchers have to interpret a non-significant result as evidence for no effect (see problem 1), face the problem that a priori power analyses are difficult (problem 3), and are required to adopt a strategy that they do not apply to other kinds of statistical analyses (problem 6). In sum, conducting an additional residual test after a significant planned contrast does not provide valuable information but leads to several problems. Given that there are alternatives that avoid at least some of the described problems, researchers could easily replace the C+R approach by one of the alternatives.

The single contrast approach avoids the problems that are associated with the residual test (Problem 3, 4, 5, and 6). Furthermore, it has the advantage that it constitutes a standard method for the comparison of group means that is discussed in many textbooks (e.g., Hays, 1988; Maxwell & Delaney, 2004; Winer et al., 1991) and statistics courses. Correspondingly, reviewers and editors are familiar with this method and it is unlikely that researchers will encounter problems when submitting manuscripts that present single contrasts. However, the single contrast approach suffers from the same *p* value related problems as the C+R approach. The outcome of the contrast test depends on sample size (problem 2) and researchers have to interpret a non-significant *p* value as evidence for no effect (problem 1) if they test predictions of no differences among group means.

Like the single contrast approach, Bayes factors and likelihood ratios avoid the problems of the residual test (problem 3, 4, and 6). Additionally, Bayes factors and likelihood ratios enable researchers to provide relative evidence for a hypothesis of no effect and, consequently, solve problem 1. In contrast to the single contrast approach, Bayes factors and likelihood ratios enable researchers to compare different models or hypotheses. Their focus on the comparison of hypotheses (or models) fosters an examination of alternative theoretical accounts and decreases the probability that researchers focus on a single model neglecting alternative theoretical accounts.

However, Bayes factors and likelihood ratios also have drawbacks. First, they share with the two other methods the dependency on sample size (problem 2). Bayes factors and likelihood ratios thus do not solve the problem that researchers might aim for the sample size that maximizes the probability of finding positive evidence for their predictions. Second, Bayes factors and likelihood ratios avoid the residual test that might lead researchers to misinterpret their results as evidence that their prediction provides the best explanation of the data (problem 5). However, researchers might also erroneously interpret a high Bayes factor or a high likelihood ratio as evidence for the general superiority of their prediction. Third, many reviewers and editors are still less familiar with Bayes factors and likelihood ratios than with *p* value based tools. Trying to publish using Bayes factors or likelihood ratios might thus be more challenging than publishing using *p* values.

In sum, the single contrast approach, Bayes factors, and likelihood ratios provide good alternatives to the C+R approach. Researchers who would like to minimize the risk of a conflict with editors and reviewers or who are interested in controlling the long-term error rates of their decisions might go for the single contrast approach.

Researchers who are interested in comparing different models or in avoiding the problems of *p* value based hypothesis testing should go for Bayes factors and likelihood ratios. In any case, the residual variance test suggested by the C+R approach should be avoided in the context of a contrast analysis.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Abelson, R. P. & Prentice, D. A. (1997). Contrast tests of interaction hypotheses. *Psychological Methods*, *2*, 315–328. doi:10.1037//1082-989X.2.4.315

Aberson, C. (2002). Interpreting null results: Improving presentation and conclusion with confidence intervals. *Journal of Articles in Support of the Null Hypothesis*, *1*, 36–42.

Blume, J. D. (2002). Tutorial in biostatistics: likelihood methods for measuring statistical evidence. *Statistics in Medicine*, *21*, 2563–2599. doi:10.1002/sim.1216

Brauer, M. & McClelland, G. (2005). L'utilisation des contrastes dans l'analyse des données : Comment tester les hypothèses spécifiques dans la recherche en psychologie ? [The use of contrasts in data analysis: How to test specific hypotheses in psychological research]. *L'année psychologique*, *105*, 273–305. doi:10.3406/psy.2005.29696

Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 189–202.

Furr, R. M. (2004). Interpreting effect sizes in contrast analysis. *Understanding Statistics*, *3*, 1–25.

Furr, R. M. & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics*, *2*, 45–67.

Glover, S. & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*(5), 791–806.

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, *130*, 1005–1013.

Hays, W. L. (1988). *Statistics* (4th ed.). New York, NY: Holt, Rinehart and Winston.

Hung, H. M., O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The behavior of the P-value when the alternative hypothesis is true. *Biometrics*, *53*, 11–22.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Johansson, T. (2011). Hail the impossible: p-values, evidence, and likelihhod. *Scandinavian Journal of Psychology*, *52*, 113–125. doi:10.1111/j.1467-9450.2010.00852.x

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis. *Behavior Research Methods*, *43*, 679–690. doi:10.3758/s13428-010-0049-5

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective* (2nd). New York, NY: Psychology Press.

Myers, J. L. & Well, A. D. (1995). *Research design and statistical analysis.* Hillsdale, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301. doi:10.1037//1082-989X.5.2.241

Niedenthal, P. M., Brauer, M., Robin, L., & Innes-Ker, A. H. (2002). Adult attachment and the perception of facial expression of emotion. *Journal of Personality and Social Psychology*, *82*, 419–433. doi:10.1037//0022-3514.82.3.419

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Rosenthal, R. & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* New York, NY: Cambridge University Press.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. doi:10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–37. doi:10.3758/PBR.16.2.225

Royall, R. M. (1997). *Statistical evidence. a likelihood paradigm.* London, UK: Chapman and Hall.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 63–71.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd). New York, NY: McGraw-Hill.

Table 1

Cell means and standard deviations (in parentheses) of Dr. Mustermann's studies.

Exam Difficulty	Study 1	Study 2
Easy	4.75 (0.72)	4.82 (0.64)
Moderate	6.75 (0.72)	5.12 (0.64)
Difficult	7.75 (0.72)	5.62 (0.64)
Very Difficult	7.85 (0.72)	5.22 (0.64)

Note. $n = 12$ in each cell. Raw data in the easy exam condition of Study 1 were 3.5, 4.0, 4.0, 4.5, 4.5, 4.5, 5.0, 5.0, 5.0, 5.5, 5.5, and 6.0. The data of the other conditions of Study 1 were created by adding 2.0, 3.0, or 3.1 to these values. Raw data in the easy exam condition of Study 2 were 4.0, 4.0, 4.2, 4.4, 4.5, 4.8, 4.8, 5.0, 5.0, 5.5, 5.6, and 6.0. The data of the other three conditions of Study 2 were created by adding 0.3, 0.8, or 0.4 to these values.

Table 2

Relevance of the six presented problems for the C+R approach, the single contrast approach, Bayes factors, and likelihood ratios.

Problem	C+R	single contrast	BF/LR
Interpreting a non-significant result as evidence for no effect (problem 1)	*	*	–
Dependency on sample size (problem 2)	*	*	*
Meaning of type I and type II errors varies (problem 3)	*	–	–
Residual may lack a meaningful interpretation (problem 4)	*	–	–
Vulnerability to misinterpretations (problem 5)	*	–	*
Underlying logic is not applied in other statistical procedures (problem 6)	*	–	–

Note. BF/LR = Bayes factors and likelihood ratios. * problem applies. – problem does not apply.