Psychon Bull Rev (2016) 23:1779–1786
DOI 10.3758/s13423-016-1026-5

BRIEF REPORT

Model comparison in ANOVA

Jeffrey N. Rouder[1] · Christopher R. Engelhardt[1,4] · Simon McCabe[2] · Richard D. Morey[3]

Published online: 11 April 2016

Abstract Analysis of variance (ANOVA), the workhorse analysis of experimental designs, consists of F-tests of main effects and interactions. Yet, testing, including traditional ANOVA, has been recently critiqued on a number of theoretical and practical grounds. In light of these critiques, model comparison and model selection serve as an attractive alternative. Model comparison differs from testing in that one can support a null or nested model vis-a-vis a more general alternative by penalizing more flexible models. We argue this ability to support more simple models allows for more nuanced theoretical conclusions than provided by traditional ANOVA F-tests. We provide a model comparison strategy and show how ANOVA models may be reparameterized to better address substantive questions in data analysis.

Keywords ANOVA · Statistical models · Interactions · Model comparison · Order-restricted inference

Factorial designs and the associated analysis of variance (ANOVA) and t-tests are workhorses of experimental psychology. Indeed, it is hard to overstate the popularity and usefulness of these designs and analyses. Consequently, most experimental psychologists are exceedingly familiar with these analyses. One would think given this popularity and familiarity that there is little room for additional development of the models and techniques underlying ANOVA. Take two-way ANOVA as an example with factors of A and B: The conventional approach is to perform three tests in this case—a main effects test of A, B, and an interaction test of A and B. How much more is there to say?

In the recent decades, statisticians and psychologists have developed methods of model comparison that go beyond traditional significance testing. In the model-comparison perspective, models instantiate theoretical positions of interest. If the models are judicious—that is, they capture theoretically important constraints—then model comparison becomes a proxy for theory comparison.

Model comparison is often similar to testing, but there is a key difference: Testing is asymmetric. One may reject a nested or null model in favor of a more general model. But the reverse does not hold—one cannot reject a more general model for a nested one. Model comparison, in contrast, has no such asymmetry: Evidence for nested or general models may be quantified. This ability to quantify evidence for nested models changes inference in ANOVA. In the two-way case, for example, instead of 3 tests, there are 8 different possible models formed by presence and absence of each main effect and the interaction. Among these 8 there are 28 possible dyadic model comparisons. The conventional tests encompass three of the possible 28. What about the others? We show here that understanding the full set of models and their relations may lead to a more nuanced understanding of the structure in data than may be provided by the more conventional tests.

Before proceeding, it is helpful to provide context on modeling itself. When one performs ANOVA, the various models formed by the inclusion and exclusion of factors

✉ Jeffrey N. Rouder
rouderj@missouri.edu

[1] University of Missouri, Columbia, MO 65211, USA

[2] University of Stirling, Stirling, FK9 4LA, UK

[3] Cardiff University, Cardiff, CF10 3XQ, UK

[4] CARFAX, inc., Columbia, MO 65201, USA

and their interactions presumably represent various positions of theoretical importance, and the differences between the models represent critical theoretical differences. If the models are good instantiations of the theoretical positions, then the inference from the models applies to the theoretical positions. This correspondence means that analysts should judiciously choose models that are theoretically interpretable. In this paper we show that these choices are

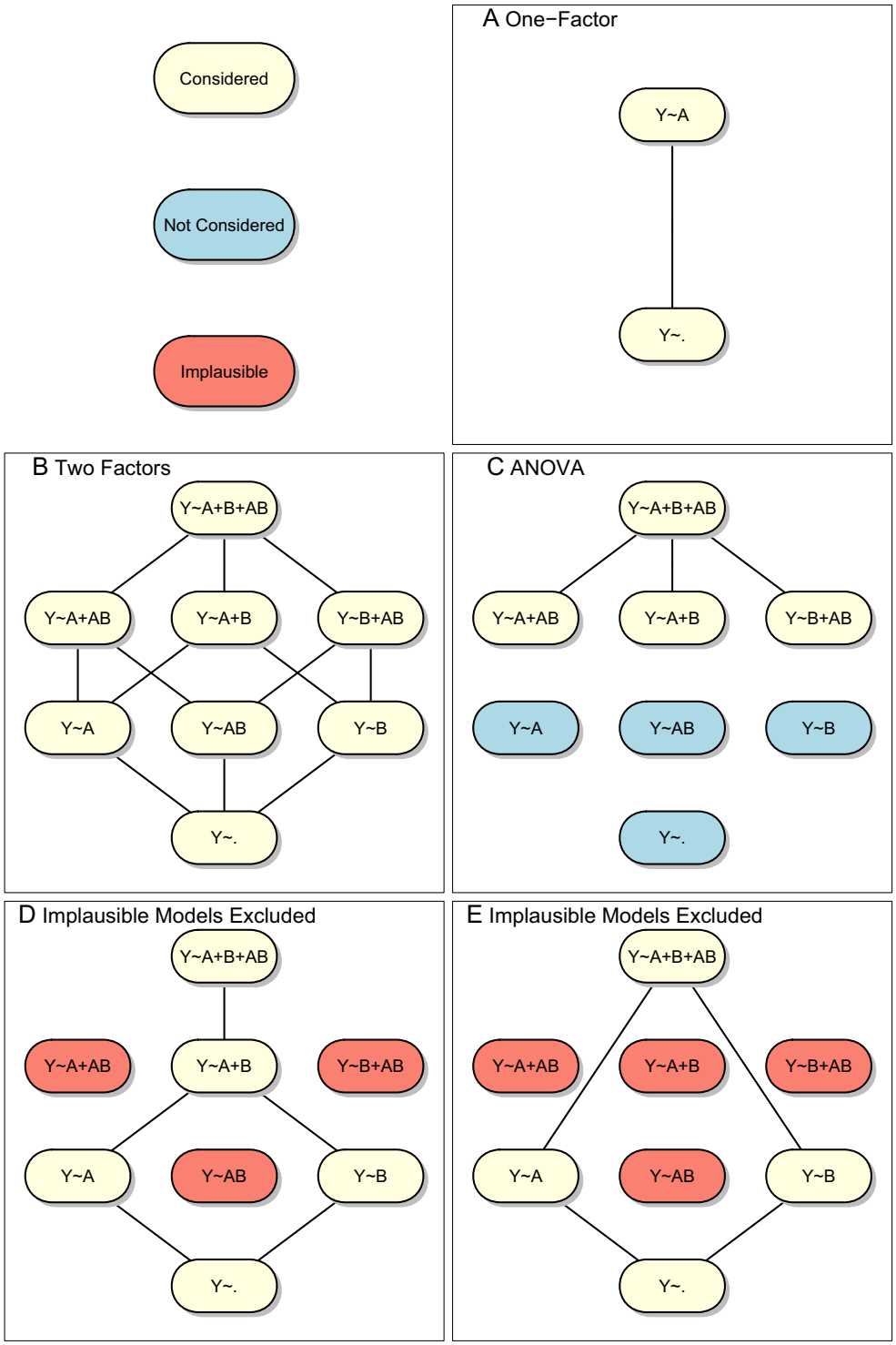

Fig. 1 Various model comparison strategies for ANOVA. **a** A comparison between a null model and an effects model for one-way ANOVA. **b** There are eight possible models for the two-way case. The lines denote nesting relations among the models. **c** Conventional ANOVA is a top-down approach that does not use the bottom of the hierarchy. **d** The exclusion of implausible models that make an exact-balancing assumption. **e** An additional exclusion when the dependent measure is ordinal rather than interval

not so simple, and, perhaps surprisingly, common models and parameterizations in ANOVA may not be the best choices for assessing main effects and interactions. In the next section, we consider one-way designs where there is a single question: did the factor have an effect. In this case, models underlying testing and model comparison are the same. Following this, we consider the two-factor case. The models for testing and comparison diverge because the ones used in testing do not, in our opinion, correspond well to the theoretical questions typically asked.

The models in a one-way design

Consider a simple one-factor design where a factor A is manipulated through some number of levels. For example, we may assess the effect of the color of presented memoranda in a memory experiment. There are two models to be considered: an *effects model* and a *null model*. The effects model may be informally written as

$$Y = \mu + A + \text{noise}.$$

This informal equation may be read as, "the response Y arises from the sum of a grand mean, the effect of the factor A, and random noise." More formal notation is possible, but it is not needed here.[1] This model may be contrasted to a null model:

$$Y = \mu + \text{noise}.$$

These two models are shown in Fig. 1A (top right) as two ovals. In the figure, the grand mean and noise terms are left out to reduce clutter and focus on the critical difference between the models, the inclusion or exclusion of factor A. The comparison of these models serves as an assessment of the necessity of factor A in describing the data. We may also refer to the null model in this context as instantiating the constraint that the data are invariant to factor A.

Model-comparison methods

In traditional tests, a dichotomous decision is made—either the analyst rejects the null or fails to do so. The failure to reject the null is not the same as accepting it, which is not formally permitted. In model comparison, in contrast, constrained models may be favored over a more general alternative. The key is to penalize more general models by their flexibility. Suppose we have two models that both account for the data. The first model can only account for the observed data and no other patterns. The second model can account for the observed data and many other patterns as well. The second model is more flexible because it can account for more data patterns. Because the more flexible model can account for a wider variety of data patterns, observing any one pattern is not so evidential. This consideration, where models are penalized by their flexibility is often called *Occam's razor*, and it is a desirable rule-of-thumb (Myung & Pitt, 1997). Using Occam's razor, we should favor the simpler constrained model which yields a constrained prediction over the alternative model which is compatible with more varied data patterns.

Our goal here is to show how consideration of models and model comparison changes how we may view ANOVA. There are several different approaches for balancing fit and flexibility in model comparison including the Akaike information criterion (AIC, Akaike, 1974), the "corrected" Akaike information criterion (AICc, Burnham & Anderson 2002), the Bayesian information criterion (BIC, Schwartz, 1978), the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & Linde, 2002), minimum description length measures (MDL Grunwald, Myung, & Pitt, 2005), and the Bayes factor (Jeffreys, 1961). Even though there are several different approaches, the points we make about models and model comparisons holds regardless of which model-comparison method is used. As an aside, we advocate for the Bayes factor (Rouder et al., 2009) because it provides for a formal, rational assessment of evidence for models. This advocacy, however, is tangential to the points provided here which hold even if one does not appreciate the Bayes factor.

Models for two-way designs

The advantages of a model comparison perspective are seen in consideration of the two-way design. A bit of terminology is helpful. The term *factor* refer to a manipulation; there are two factors in a two-way design. The term *covariate* refers to statistical variable. There are three covariates in a two-way design: the main effect of A, the main effect of B, and of the AB interaction. Each covariate may contain a number of parameters depending on the number of levels.[2] The full ANOVA model for this design is stated informally as

$$Y = \mu + A + B + AB + \text{noise},$$

[1] The grand mean μ is a single free parameter. The term A denotes the effects of Factor A, and it is a vector of parameters with one parameter for each level of the factor. If these levels are assumed to be fixed effects, then a sums-to-zero constraint may be placed on the parameters yielding one fewer parameters than levels.

[2] Let a and b denote the levels of factors A and B, respectively. Then, there are a, b, and $a \times b$ parameters for the main effects of factors A and B and the interaction, respectively. If A and B are treated as fixed effects, then the imposition of the usual sums-to-zero constraints yields $a-1$, $b-1$ and $(a-1)(b-1)$ parameters for the main effects of factors A and B and the interaction, respectively.

This full model occupies the top of a model hierarchy shown in Fig. 1B. The bottom of the hierarchy is a null model. Above the null model are the three models with a single covariate: a model with the the main effect of A and no other effects, a model with the main effect of B and no other effects, and a model with the AB interaction and no other effects. Above these three models are an additional three models with any two of the three covariates. The lines in the figure show the nesting relations among models.

The conventional ANOVA testing approach is shown in Fig. 1C. It is a top-down strategy where the full model is compared to models without a specific covariate. Evidence for the missing covariate is inferred when the full model is preferred to the model missing that covariate. For example, there is evidence for A when the full model is preferred to the submodel $Y = \mu + B + AB +$ noise that excludes A.

The advantages of considering many models

Our main claim is that by considering all plausible models, the analyst can make more theoretically detailed statements about the structure of data. This claim is perhaps best illustrated by consideration of models that are not part of the ANOVA testing approach. Take for example, the model $Y = \mu + A +$ noise. This model is comprised by an effect of A and an invariance to B. These are both theoretically appealing statements. Suppose Y is a free-recall score to memorized words; A is the part-of-speech of the words, say noun vs. verb, and B is the color of the words. These factors were chosen in fact to assess whether the recollection process was driven by semantic information, in which case part-of-speech should affect free recall, or by perceptual information, in which case color should affect free recall. The model instantiates the case that semantic information rather than perceptual information drives recollection. It is a theoretically useful and important model.

In model comparison approaches, a numeric value is placed on each model in the hierarchy. For example, if AIC was the target comparison statistic, an AIC value would be computed for each model. Suppose the model $Y = \mu + A +$ noise is favored over all other models. This model is immediately interpretable: The evidence for the effect of A is in the comparison to the null; the evidence for the invariance to B is in the comparison of $Y = \mu + A + B +$ noise. This approach may be contrasted to a top-down testing approach where the comparable statements are made when the full model fits significantly better than the model missing the main effect of A but not significantly better than models with $A+B$ or $A+AB$. The model-comparison and top-down testing approaches will assuredly differ if for no other reason that it is impossible to interpret the lack of significance as evidence for an invariance. By adhering to Occam's razor and allowing preference for nested models, the model comparison approach offers a richer, more insightful view of the structure in data.

Implausible models

Figure 1B shows eight models, but not all of them are theoretically interpretable. We show here that several should be discarded *a priori* in most cases because they correspond to positions that are *a priori* implausible in most experimental settings. Statisticians routinely voice concern about testing main effects in the presence of interactions because the size and sign of the main effects depend on the levels of the factors (Nelder, 1998; Venables, 2000). We have a related concern about models with interactions but without corresponding main effects. These models are implausible

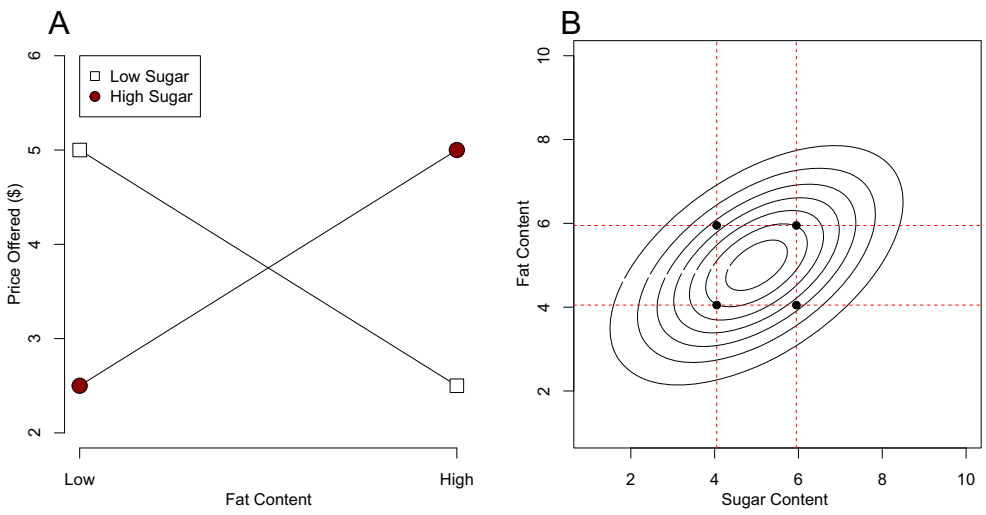

Fig. 2 **a** An interaction without main effects for the price offered for a pint of ice cream as a function of sugar and fat content. **b** The same dependency across a continuum of levels. From this panel it is clear that the lack of main effects reflects an implausibly fortuitous choice of levels

because they seemingly rely on picking the exact levels so that the true main effects perfectly cancel. To illustrate this claim, consider an experiment to explore how much people would be willing to pay for a pint of vanilla ice-cream as a function of its sugar and fat content. Figure 2A shows a hypothetical example. Here, the sugar and fat need to be matched to maintain balance otherwise the ice-cream tastes too sweet or too bland. As such, perhaps the lack of main effects may seem plausible. To show that it is not, consider Fig. 2B which shows a more complete view of the relationship between sugar content, fat content, and value of the pint of ice-cream. The inner most circle shows the highest value, in this case above $6 a pint, and the successive rings show lessening value. The balance comes from the positive association, for each level of fat there is a balancing level of sugar. The points, at values of $2.5 and $5, correspond to the points in Fig. 2a. For the shown levels there are no main effects. But, if the levels are chosen differently, then main effects surely appear. That is, the lack of main effects here results only for very carefully and precisely chosen levels, and main effects appear otherwise. In application, we do not have the information in Fig. 2b (otherwise, we would not had to perform the experiment), and it seems implausible that one could fortuitously choose levels to precisely cancel out the main effects.

Because we find models with perfectly cancelling level to be implausible, we discard all models where there is an interaction without corresponding main effects. In the two-factor case, the discarded models are shown in red in Fig. 1d. Consider $Y = \mu + B + AB + \text{noise}$. It contains an interaction involving A, the AB interaction, without a specification of the main effect of A. The two other discarded models are $Y = \mu + A + AB + \text{noise}$, which contains an AB interaction without a main effect of B and $Y = \mu + AB + \text{noise}$, which is missing main effects of both variables.

The consequence of these discards is that it takes stronger patterns in data to evidence interactions. Interactions occur only in the presence of main effects, and models with main effects are more flexible than those without them. And with this increased flexibility comes an increased penalty. Such a consequence is desirable. Evidence for interactions should not depend on assuming the levels are perfectly chosen so that main effects cancel.

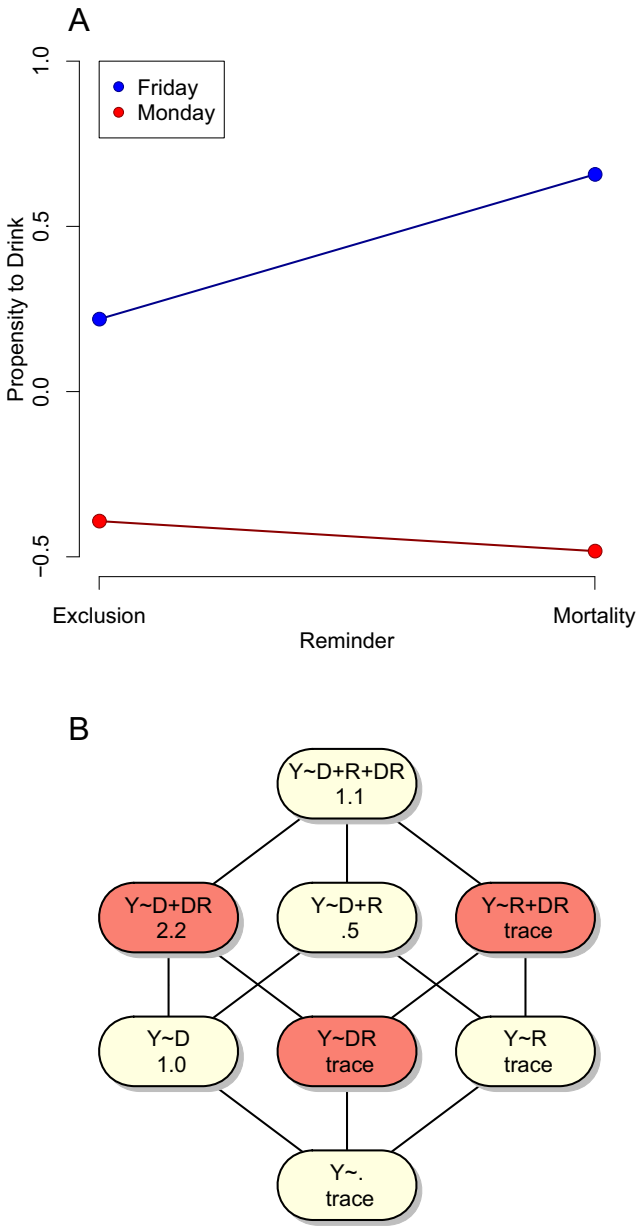

Fig. 3 Results from McCable et al.: **a** Sample means indicate the plausibility of a main effect of day and an interaction between day and reminder type. Error bars denote standard errors. **b** Bayes factor model comparisons for the standard ANOVA parameterization. The label "trace" refers to very small Bayes factor values on the order of 10^{-10}. These values hold for models without a main effect of day. The models colored in light *yellow* are considered while those colored in *red* are discarded as implausible because they assume interactions that perfectly cancel out main effects

Plausible models for ordinal dependent measures

Experimental psychologists use a wide variety of dependent measures. Some of these dependent measures are on a ratio or interval scale, and examples include response times and voltages. Other dependent measures are on an ordinal scale where differences are not interpretable as a metric quantity. An example of an ordinal measure includes confidence ratings scales. Here differences between two ratings, say between "1" and "2" on the scale, cannot be assumed to be the same across the scale. For example the difference between "1" and "2" may be psychologically different than that between "2" and "3." When differences between such intervals are not constant, additivity is not a meaningful concept, and models without interactions are difficult to

Table 1 The usual parameterization of the 2-by-2 ANOVA model

	Exclusion	Mortality	Mean
Monday	$\mu - \alpha - \beta + \gamma$	$\mu - \alpha + \beta - \gamma$	$\mu - \alpha$
Friday	$\mu + \alpha - \beta - \gamma$	$\mu + \alpha + \beta + \gamma$	$\mu + \alpha$
Mean	$\mu - \beta$	$\mu + \beta$	μ

interpret because the constraint is unjustifiable. As a consequence, the effect of a factor is best assessed by considering whether it enters in full or not at all.

Experimental psychologists often consider proportions, say the proportion of one type of response or another, as the dependent measure of interest. There are hundreds of studies that test interactions among factors for this measure. By our reckoning, these types of contrasts need additional care because the meaning of a difference in a proportion is dependent on the value of the proportion itself. For example, small differences for small valued proportions, say those around .05 have a different meaning than those around moderately valued ones, say around .5. In this context, it is difficult to interpret the contrast between a main effects model and a full model, because the meaning of the main effects model is unclear. Hence, in our opinion, claims of interactions should be avoided. Instead, researchers should limit themselves to describing whether a factor has an effect without discriminating between main effects and interactions. To implement this limit, we do not use main-effects models in our model-comparison analyses when using ordinal data. This additional discard is shown in Fig. 1e.

An example of model comparison

To illustrate the model comparison approach we take an example from McCabe et al. (submitted). Their goal was to assess the effect of reminding people of their own mortality on their attitudes toward alcohol consumption. The control condition was the attitudes when asked to consider a situation where they felt socially excluded, and the difference in attitudes served as a measure of effect. The experiment was motivated by a psychological theory, terror management theory (Greenberg et al., 1986), that describes how our recognition of our mortality affects our thoughts, behavior, and actions. According to this theory, thoughts about death causes people to cling more tightly to social norms. Drinking alcohol is part of the normative cultural script for Fridays (a leisure day) but not for Mondays (a work day). Consequently, McCabe et al. hypothesized that the mortality reminder might produce more positive attitudes toward drinking alcohol on Fridays than on Mondays. An interaction effect between the day of assessment and the presence of the a morality reminder was hypothesized.

Cell means are plotted in Fig. 3a. As can be seen, there seems to be a healthy interaction, and indeed, the conventional statistics reveal no main effect of the mortality reminder ($F(1, 156) = 2.28$, $p \approx .13$), an unmistakable main effect of day ($F(1, 56) = 57.9$, $p \approx 10^{-11}$), and an interaction between the two ($F(1, 56) = 5.28$, $p \approx .023$). By conventional interpretation, the interaction seems supported.

We illustrate the effect of discarding implausible models with model comparison by Bayes factors. The points we make are about the strategy of model comparisons rather than about Bayes factors, and they hold broadly across model comparison measures that appropriately penalize complexity. The Bayes factor model comparison statistics are shown in Fig. 3b.[3] We took the model $Y = \mu + D + \text{noise}$, on the lower left, as the standard and compared all models to it. Consequently, it has a value of 1.0 in the figure. Models without the main effect of day fared poorly in comparison; the Bayes factor values for these models is on the order of 10^{-10} (one in 10 billion), which is denoted in the figure as "trace" values. If we temporarily (and wrongly) consider all models, the best model is $Y = \mu + D + DR + \text{noise}$, which has a Bayes-factor value of 2.2 compared to the standard model $Y = \mu + D + \text{noise}$. The interpretation here is that the data are about twice as likely under the model with the interaction than without, and that Bayesians may update their odds in favor of the interaction by a factor of 2.2.

The above interpretation, however, rests on a perfect-balance assumption: the upward effect of the mortality cue presented on Friday is balanced exactly by mirror downward effect of the same cue presented on Monday. To dispense with this unwarranted assumption, we compare the Bayes factor of the most favored interpretable model *with* an interaction, $Y = \mu + D + R + DR + \text{noise}$ to the most favored model *without* it, $Y = \mu + D$. This Bayes factor is 1.07 in value, which is equivocal. What happened is straightforward: The interpretable interaction model, $Y = \mu + D + R + DR + \text{noise}$, is more flexible than the discarded model, $Y = \mu + D + DR + \text{noise}$. Hence, it is penalized more and this increased penalty accounts for the attenuation of evidence.

[3] Bayes factors described in Rouder et al. (2012) were computed with the anovaBF function in the BayesFactor package for R (Morey and Rouder, 2015). Default values of .5 on the scale of effect sizes were used.

Table 2 The cornerstone parameterization of the 2-by-2 ANOVA model

	Exclusion	Mortality	Mean
Monday	μ	$\mu+\alpha$	$\mu+\frac{\alpha}{2}$
Friday	$\mu+\beta$	$\mu+\alpha+\beta+\gamma$	$\mu+\frac{\alpha}{2}+\beta+\frac{\gamma}{2}$
Mean	$\mu+\frac{\beta}{2}$	$\mu+\alpha+\frac{\beta}{2}+\frac{\gamma}{2}$	$\mu+\frac{\alpha}{2}+\frac{\beta}{2}+\frac{\gamma}{4}$

Reparameterizing the ANOVA model

In the preceding development, we showed how models with interactions but without main effects are implausible, and we provided an example where discarding them had a substantive effect on the conclusion. The critical problem is the perfect-balance assumption. Here we construct a model that captures the notion of an interaction without being too flexible and without making the perfect balance assumption. To arrive at this model, we first explore the conventional parameterization of the ANOVA model. Then we reparameterize so that main effects have subtly different meaning. With this alternative parameterization it is possible to have interactions without main effects while not committing to perfect balance.

The usual parameterization of ANOVA is shown in Table 1, and it is presented in the context of the McCabe example for concreteness. The grand mean is denoted by μ, the main effect of day is denoted by α, the main effect of the reminder is denoted by β, and the interaction is denoted by γ. Note the symmetry that each effect is present in all cells. The alternative parameterization, called the *cornerstone parameterization*, is shown in Table 2. In this parameterization, one condition—in our case the social-exclusion-reminder-on-Monday condition—serves as a baseline or cornerstone. All effects are measured relative to it.

Here, μ serves as a baseline parameter rather than a grand mean. Although α and β are main effects, they are measured relative to baseline rather than to the grand mean. And γ, the interaction, is the deviation from additivity as manifest in a single cell.

Parameters in the usual and cornerstone parameterizations have different meanings. The cornerstone parameterization appears inconvenient because main effects are relative to only one condition rather than marginal across several conditions. Moreover, the marginal means across rows and columns include all parameters. The key benefit of this parameterization is that interactions may occur without main effects and, importantly, without a perfect balance assumption. If both main effects are zero, the resulting data pattern is that three cells have the same value while the fourth has a different one. Such a pattern is plausible; for example, it would describe plant growth where one factor is whether water is provided or not and the other factor is whether sunlight was provided or night. Plants grow only with some sunlight and water. Importantly, there is no perfectly opposing balance conditions when the main effects are zero.

The cornerstone parameterizations strike us as particularly appropriate when the theoretical question of interest is limited to the interaction because the main effects describe only the effect at one level. It is well-suited to McCabe et al.'s analysis because in this study there is no specific theoretical question about the main effects of day or type of reminder. Here, the theoretical question is about the over-additive interaction, or a positive value of γ. The assessment of γ may be made regardless of the values of α and β. To assess the McCabe et al. claim, we computed the Bayes factors between a model with $\gamma = 0$ (additive effects) and one with $\gamma > 0$ (over additive interaction).[4] The Bayes factor is 9.2 in favor of the interaction, indicating evidence for the hypothesis that mortality reminders lead to more positive attitudes toward drinking alcohol on Fridays than on Mondays. This Bayes factor, from a more targeted and thoughtful modeling implementation, best captures the theoretically important structure in the data.

Summary

From the model-comparison perspective, positions of theoretical importance are instantiated as models. These models are then compared and the evidence from the comparison is a proxy for evidence for competing theoretical positions. The benefit of modeling, as is shown here, is that one can tailor the model precisely to match the theories under consideration, providing inference with more resolution than is possible with the off-the-shelf ANOVA procedures. Judicious modeling is assuredly a path to a more fruitful, insightful, accurate, and productive psychological science.

Author Note Email: rouderj@missouri.edu, Web: pcl.missouri.edu; Twitter: @JeffRouder. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

[4]Bayes factors were computed with the nWayAOV function in the BayesFactor package. This function allows the analyst to set the design matrix, which was constructed to encode the alternative parameterization. Default values of .5 on the scale of effect sizes were used.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information theoretic approach*, 2nd edn. New York: Springer.

Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In Baumeister, R.F. (Ed.) *Public self and private self*. New York: Springer.

Grunwald, P., Myung, I.J., & Pitt, M.A. (2005). *Advances in minimum description length: Theory and applications*. Cambridge: MIT Press.

Jeffreys, H. (1961). *Theory of probability*, 3rd edn. New York: Oxford University Press.

McCabe, S., Arndt, J., Bartholow, B.D., & Engelhardt, C.R. (submitted). Mortality salience and the weekend effect: Thoughts of death enhance alcohol related attitudes on Friday but not on Monday.

Morey, R.D., & Rouder, J.N. (2015). BayesFactor 0.9.12-2. Comprehensive R Archive Network. Available from http://cran.r-project.org/web/packages/BayesFactor/index.html

Myung, I.-J., & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Nelder, J.A. (1998). The selection of terms in response-surface models-How strong is the weak-heredity principle? *American Statistician*, *52*, 315–318.

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. Available from 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 and Review*, *16*, 225–237. Available from doi:10.3758/PBR.16.2.225

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *64*, 583–639.

Venables, W.N. (2000). *Exegeses on linear models. Paper presented to the S-PLUS User's Conference. Available from,* http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf