Computational Brain & Behavior (2023) 6:28–34
https://doi.org/10.1007/s42113-021-00120-3

ORIGINAL PAPER

The Limits of Marginality

Andrew Heathcote[1] · **Dora Matzke**[2]

Accepted: 6 September 2021 / Published online: 22 September 2021

Abstract
The "marginality principle" for linear regression models states that when a higher order term is included, its constituent terms must also be included. The target article relies on this principle for the fixed-effects part of linear mixed models of ANOVA designs and considers the implication that if extended to combined fixed-and-random-effects models, model selection tests specific to some fixed-effects ANOVA terms are not possible. We review the basis for this principle for fixed-effects models and delineate its limits. We then consider its extension to combined fixed-and-random-effects models. We conclude that we have been unable to find in the literature, including the target article, and have ourselves been unable to construct any satisfactory argument against the use of incomplete ANOVA models. The only basis we could find requires one to assume that it is not possible to test point-null hypotheses, something we disagree with, and which we believe is incompatible with the Bayesian model-selection methods that are the basis of the target article.

Keywords Linear mixed models · Response surface models · ANOVA models

In their paper, van Doorn et al., (2021; hereafter vDAHSW) discuss how to conduct model selection with linear mixed models for ANOVA designs using Bayes Factors (Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012). In this comment, we examine their suggestion that the idea of "marginality" might be used to rule out some models a priori. The exclusion only allows models with an interaction or interactions when they also include all of the components of the interaction(s). Following the earliest paper, we know of on this topic (Bernhardt & Jung, 1979), we will call the allowed models "complete".[1] vDAHSW make this restriction on the fixed-effects part of a mixed model, based on a fixed-effects example provided by Wagenmakers et al. (2018) and references cited therein. They also consider generalizing it to the combined fixed-and-random-effects model, citing Rouder et al. (2016), which was the original source of the fixed-effects example (reproduced also in Rouder et al., 2017).

We will first address the fixed-effects argument, tracing the detailed history of this issue in the statistical literature. This literature addresses only the case of regression models with continuous predictors (i.e., independent variables), making arguments that rely on the ratio-scale properties of those predictors. vDAHSW cite much of this literature and appear to use it to support a generalization to ANOVA analyses, yet such analyses do not rely on independent variables having a ratio-scale property, or even refer to this property in any way. We believe that our detailed review of the continuous case shows that its relevance to vDAHSW's ANOVAs is far from obvious, so needs to be justified. We then turn to Rouder et al's. (2016) fixed-effects example, showing that if it is treated as a continuous regression problem, it is in fact a case where incomplete models are explicitly sanctioned as acceptable in the statistical literature. We then show that the specific argument based on this example made by Rouder et al. against incomplete models is unrelated to the issues raised in the statistical literature, and it is

✉ Andrew Heathcote
andrew.heathcote@newcastle.edu.au

[1] School of Psychology, The University of Newcastle, Callaghan, Australia

[2] Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

[1] vDAHSW's Model 6 is an example of a complete model, and their Model 5 an example of an incomplete model. The terms "well-formulated" (Peixoto, 1987, 1990) and "well-formed" (Nelder, 1998) have also been used for completeness, but we avoid these because of the potential implication of a value judgement which we will argue that best applies to a restricted set of models with particular sorts of continuous independent variables. The usage "marginality" appears to be based on Nelder's (2000) use of "functional marginality," which is again restricted to these particular sorts of independent variables in the context of a particular type of continuous regression analysis, response-surface modeling, which we discuss in more detail later.

tantamount to asserting that a point-null hypothesis cannot be proved, something that goes against one of the most often cited advantages of the Bayesian model-selection approach endorsed by vDAHSW. We finish by discussing a key question for vDAHSW: Under what circumstances (if any) is it valid to entertain a combined fixed-and-random-effects model that is incomplete? If completeness is required for such mixed effects, then model selection cannot be used to test for the absence of a particular fixed effect while allowing for the presence of an interaction of that fixed effect with a random effect, severely limiting the specificity of allowable tests. We conclude that we do not know of any rigorous justification for this restriction for the case they address, with a random participant effect, or indeed for other common cases such as random item effects. We hope that this comment will, therefore, prompt vDAHSW to provide a more detailed justification for their position with respect to both incomplete fixed-effects models and incomplete mixed-effects models.

Completeness in Regression Models with Continuous Predictors

The issue of completeness was raised in the statistical literature in the context of regression models with continuous predictors that are not on a ratio scale. The critical property of a ratio-scaled predictor in this context is that it has a zero point. Using an example after Nelder (1998), consider the regression equation[2]:

$$Z = a + bY \tag{1}$$

If Y is not ratio scaled, we might, with equal validity, use $W = m + Y$ as a predictor for some arbitrarily chosen m on the real line, and hence the regression equation becomes

$$Z = a' + bW, \tag{2}$$

where $a' = a - bm$. It makes no sense to select between a model with and without an intercept term, or in any other way test the intercept term, because the presence or absence of an intercept depends entirely on our arbitrary choice of m. That is, the estimated intercept is not interpretable because the arbitrary nature of the zero point of an interval scale makes the intercept's estimated value arbitrary. Similarly, selection between a model with and without an estimated intercept is not interpretable because the difference between these two models is arbitrary (i.e., if $m = a/b$ then $Z = bW$, and so the intercept disappears). This is the reason that statisticians often caution against the use of models without an intercept, and against model selection or other tests involving the intercept term. Importantly, however, models without an intercept are entirely interpretable when using ratio-scaled predictors unless their natural scale is not respected (e.g., they are centered or standardized).

Bernhardt and Jung (1979) first addressed issues related to completeness from a more general perspective, for both a regression with two predictors that interact, and for a single predictor in a polynomial regression. We will focus on the former case, as it is most germane to vDAHSW. Consider the regression:

$$Z = a + bX + cC + dXC \tag{3}$$

And an inhomogeneous scaling transformation $F = m + n\mathrm{C}$. An example of the latter is converting from the Celsius scale to the Fahrenheit scale for a temperature predictor, in which case $m = 32$ and $n = 9/5$ (Griepentrog et al., 1982). As before, m is on the real line, and n is positive (a necessary condition for this transformation to be a proper scaling). Note that none of the completeness issues raised next arise if we enforce $m = 0$ (in which case the transformation is described as "homogenous"), or if the regression does not contain the interaction term or terms above linear in the polynomial case (except of course with respect to dropping the intercept as already discussed). It then follows that

$$Z = a' + b'X + c'F + d'XF, \tag{4}$$

where $a' = a - (m/n)c$, $b' = b - (m/n)d$, $c' = c/n$, and $d' = d/n$. Once again, it is meaningless to specify a model without an intercept (as we can always make the intercept zero by choosing $m/n = a/c$), and now it is also meaningless to specify a model without a term that is linear in X (as we can always make that term disappear by choosing $m/n = b/d$). The same argument can be made with respect to the term that is linear in C if X is not ratio scaled.

Table 1 gives two numerical examples of these algebraic facts, providing regression coefficients on the Fahrenheit scale that produce identical values of the dependent variable (Z) to the regression coefficients given for the regression on the Celsius scale. In both Examples 1 and 2, a non-zero intercept on the Celsius scale (corresponding to Eq. (3)) becomes a zero on the Fahrenheit scale (corresponding to Eq. (4)). These examples illustrate why it can be misleading to compare models with and without an intercept: On the Celsius scale, the model with the intercept would be selected, whereas on the Fahrenheit scale, the model with no intercept would be selected. In Example 2, the scale transformation causes the non-zero coefficient b in (3) to become zero in (4) (i.e., $b' = 0$), and hence any effect of X to also

[2] As noted by McCullagh and Nelder (1989), these issues apply to any generalized linear model, so Z can be discrete or continuous. For brevity, we will suppress mention of the random part when discussing the regression setting.

Table 1 Numerical examples of how regression coefficients change when rescaling the variable C in regression Eq. (3). The examples concern rescaling to Fahrenheit = 32 + 9/5 × Celsius (i.e., $m = 32$, $n = 9/5$), where the intercept for the regression using the Fahrenheit scale is zero (i.e., $a' = 0$) for any values of a and c such that $a/c = m/n = 160/9$, and for any values whatsoever of b and d. For numerical simplicity, we use $a = 160$ and $c = 9$. The examples use two pairs of values for b and d, both of which remove the intercept, and the second of which removes X as well (as $m/n = b/d$)

	Scale	Intercept	X	Temperature	Interaction
Example 1	Celsius	$a = 160$	$b = 161$	$c = 9$	$d = 9$
	Fahrenheit	$a' = 0$	$b' = 1$	$c' = 5$	$d' = 5$
Example 2	Celsius	$a = 160$	$b = 160$	$c = 9$	$d = 9$
	Fahrenheit	$a' = 0$	$b' = 0$	$c' = 5$	$d' = 5$

disappear. This illustrates that comparing models with and without the X effect can also be meaningless.

If X and C (and any number of additional predictors) are continuous and not on their natural ratio scale, only complete regression models make sense as the constituents of higher order terms (called by Peixoto, 1990, "inferior" terms), and incomplete models in general, are not interpretable. Peixoto (1990) illustrates the lack of interpretability in an example where two automatic variable selection algorithms not respecting completeness produced a puzzling inconsistency when one required a predictor to be centered for numerical stability (see Peixoto, 1987, for a variable selection method in polynomial regression models that addresses these problems by considering only complete models). However, it is important to note that these problems of interpretation do not apply to continuous predictors that are ratio scaled: "There is nothing wrong with the use of polynomials with missing inferior terms to describe exact laws of, for example, physics and chemistry" (Peixoto, 1990, p. 29–30).

We would disagree with Peixoto (1990) only in the generality of the example, as ratio-scaled predictors are certainly not the sole domain of the hard sciences. Indeed, we will shortly turn to Rouder et al.'s (2016) psychological example where the predictors are ratio scaled. Before doing so, we note a refinement of the forgoing arguments with respect to regressions involving multiplicative terms in response-surface models (i.e., a regression model that predicts a dependent variable's value at all points within a region defined with respect to the predictors). Nelder (1998) showed that some incomplete models can make sense when there is a "special point" on (say) predictor X where predictor C has no effect. The example given was a slope-ratio assay, where X is the dose of a drug and C is a synergist (i.e., a catalyst) that linearly enhances the effect of X. At the "special point" $X = 0$, the synergist C has nothing to act on, and so the response (Z) will be independent of C (although not necessarily zero). Hence, the incomplete model $Z = a + bX + dX$C remains sensible and so is admissible for consideration. The reverse is not true (i.e., $Z = a + cC + dXC$ is not interpretable), but if both X and C have such special points, then $Z = a + dXC$ is interpretable.

Completeness in ANOVA Models

At this point, we admit to being puzzled as to why vDAHSW cited some of the literature just reviewed (with the Rouder et al., 2017, and Wagenmakers et al., 2018, papers they reference citing the remainder) when their topic is ANOVA designs. Our puzzlement stems from the fact that ANOVA models do not rely on the ratio scaling of independent variables. At most, polynomial coding relies on interval scaling, and all other coding schemes such as dummy (i.e., treatment) or sum coding do not even require that. A reviewer wondered whether the use of dummy coding (i.e., the use of 0 and 1 values to convey level membership) is relevant here. In this coding, the level with all zero values, or with all -1 values in sum coding, can be thought of as providing a reference relative to which the effects of other levels are compared. However, the idea of zero point and the idea of a reference point seem to us quite different, and in these coding schemes, nothing about the scale values of any factor level is used in the ANOVA computation (unlike the issues raised in the statistical literature, where scale values are integral to the regression computation), so our puzzlement remains.

vDAHSW also cited the example first given by Rouder et al. (2016) as a basis for considering only complete fixed-effects models, and as a basis for expressing doubts about combined fixed-and-random-effects models that are incomplete (i.e., Model 5 in their Table 1). However, this example raises a quite separate potential problem from those just discussed, the issue of "perfect balance." The example is reproduced in Fig. 1, with predictors X = sugar content, Y = fat content (here we change our notation from C used in (3) and (4) to be appropriate to the Cartesian coordinates in Fig. 1), and the dependent variable is Z = the price offered for icecream. Figure 1a represents Z as a response surface (i.e., a contour plot of Z as a function of X and Y). Figure 1b shows the results of an experiment measuring just four points on the surface with low vs. high sugar and fat content factors that produces a perfect crossover interaction. This pattern corresponds to an incomplete ANOVA model with an interaction but no main effects, which vDAHSW characterize as "only plausible when the exact levels of each factor are picked such that the true main effects perfectly cancel, which in most practical applications seems implausible" (p. 10).

This example relies on three parts: (1) an underlying response-surface defined on continuous predictors; (2) choosing particular values of these predictors to construct

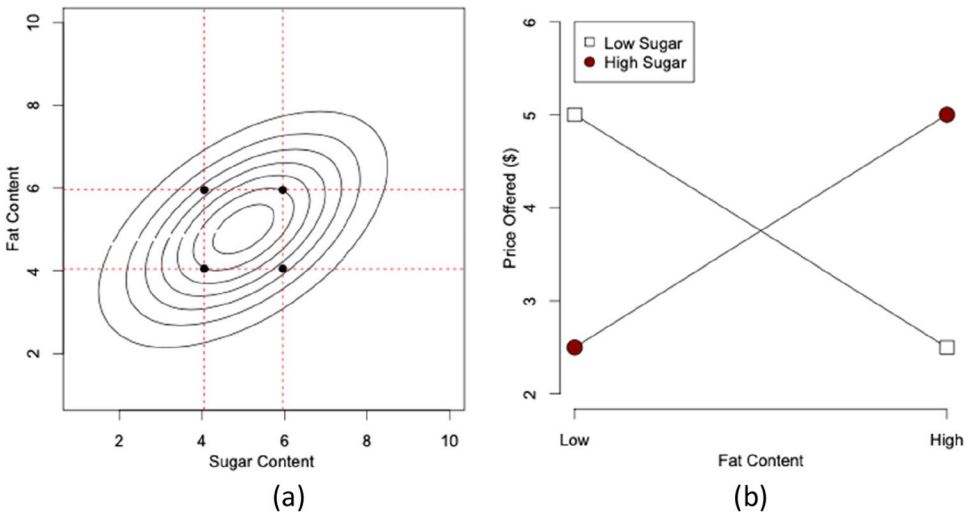

Fig. 1 Rouder et al's. (2016) icecream example. In (**a**), lower values of price offered for the outer equal-value ellipses change to higher values for the inner ellipses. (**b**) plots the height of the four points marked by dots in (**a**)

an ANOVA design constituted of factors with discrete levels; (3) and those choices resulting in perfect balance. In order not to confuse these parts, we first address the analysis of the underlying continuous response-surface model and then next the ANOVA analysis and perfect balance parts. First, because X and Y are ratio scaled—$X=0$ (no sugar) and $Y=0$ (no fat) are not arbitrary—it is an example where there is no dispute that an incomplete response-surface model is admissible. A reviewer pointed out that the mapping from objective fat and sugar content to the corresponding subjective quantities that presumably mediate taste preference, and in turn determine the amount paid, may not have known zero points. We agree that caution must be exercised with respect to inferences about subjective quantities based on analyses of objective quantities. If there is a basis to assume that the objective and subjective mappings only differ in their zero points, then ruling out incomplete models may be helpful. However, other plausible potential issues, such as the mapping being nonlinear, may still be problematic, potentially requiring either explicit cognitive models (e.g., van Ravenzwaaij et al., 2020) or in some cases specialized non-parametric analyses that rely on relatively weak assumptions such as monotonicity (e.g., state-trace analysis; Bamber, 1979; Prince et al., 2012). However, this issue strikes us as a much broader one than is being addressed here, and it does not change the fact that inferences about the effects on behavior of directly observable ratio-scaled quantities can be based on incomplete regression models with continuous ratio-scaled predictors.

With respect to the second part, constructing an ANOVA model, we certainly agree that measuring a response surface with four points is unwise and can easily lead to spurious conclusions. However, we do not see how such unwise design choices have any relationship to the status of incomplete ANOVA models. A potential clue is afforded by Venables (2000), which was cited by Wagenmakers et al. (2018) in support of the a priori exclusion of incomplete ANOVA models. Venables (2000) shares Nelder's (2000) "functional marginality" nomenclature, and although it deals with polynomial regression, it does address concepts shared with ANOVA models, as the following quote illustrates: "the intercept and linear terms are marginal to the quadratic term, a concept guaranteed to generate controversy like no other in this area. Marginality is also at the crux of another thorny issue in linear models, namely that of Type III sums of squares" (p. 9). Indeed, Venables draws a conclusion that one could certainly be forgiven in thinking generalizes to ANOVA models: "… testing main effects in the presence of an interaction is a violation of the *marginality principle*. This is not a totally rigid principle, but in all common practical situations the sensible thing is to respect it" (p. 13). Wagenmakers et al. appear to assume that Venables' conclusion about polynomial regression generalizes to ANOVA models, making the following statement (immediately followed by a list of all of the papers that we have now reviewed, i.e., papers that only deal with continuous predictors): "Consistent with the principle of marginality, JASP does not include interactions in the absence of the component main effects; for instance, the interaction-only model "Gender × Pitch" may not be entertained without also adding the two main effects" (p. 69). However, in contrast to polynomial regression, there is nothing about any continuous values that might be associated with the levels of the gender or pitch factors that in any way enter into the calculation of the ANOVA, and so we are unable to understand the relevance of the "marginality principle" to such ANOVA models.

The third part is summarized by both Rouder et al. (2016) and Rouder et al. (2017) using identical words: "the lack of main effects reflects an implausibly fortuitous choice of levels" (p. 1782 and p. 314, respectively). We agree that this example reflects a poor choice in unnecessarily dichotomizing continuous predictors. However, we wonder how this

example justifies the conclusion, to repurpose Venable's (2000) words, that in all common practical situations, the sensible thing is to respect the principle of using only complete fixed-effects ANOVA models. Unfortunately, we have been unable to generate such a justification. To illustrate our problem in doing so, we give our own example, which also provides a basis for later considering the issue of interactions between fixed effects and random-participant effects.

Consider an experiment in which males and females (factor "gender" or G) are measured on their reaction to the presence or absence of a particular stimulus or the presence of one or other of a pair of stimuli that differ in some nominal way (e.g., having a convex vs. concave shape, factor "stimulus" or S). Consider selection of a complete model, one with only a stimulus main effect where (say) the response (Z) is greater for the second level of S than the first level of S by some amount s, denoted[3] by $Z \sim S$. This selection asserts two point-null effects or "invariances" related to gender, both in the overall level of Z and in a gender difference in the magnitude of the stimulus effect. We assume this selection to be uncontroversial, at least in the present context where Bayes factors can be used to support these two point-null statements. If one were to visualize this model, it would look like Fig. 1a with stimulus on the x-axis and gender differentiating the lines, but with the order of the points on the x-axis flipped for the low-sugar line (i.e., so that both lines lie on top of each other).

Now consider an incomplete interaction-only model: $Z \sim G{:}S$. For example, the stimulus effect for males could be reversed, with a greater response for the first level of S than the second by an amount s. We cannot see why this model is so much less plausible than the S main-effect-only model that it can be ruled out a priori. Both models assert two point-null effects of the overall gender effects, and for the interaction-only model sensitivity to the stimulus effect (i.e., its absolute magnitude). They differ only in the sign of the stimulus effect for males. Given there appears to be no interpretability issue associated with measurement scales in this example, and particularly as opposite effect directions between the genders (or indeed any other fixed effects) are potentially sensible, it seems to us that to a priori rule out this model because it is incomplete, results in an inferential framework incapable of discovering psychologically interesting findings. Put another way, given Rouder et al's. (2016) view that model selection based on Bayes factors offers "a richer, more insightful view of the structure in data" (p. 1782) when it identifies a point-null main effect in the absence of an interaction, we do not see why it cannot provide those rich insights in identifying two point-null main effects when an interaction is present. In summary, Rouder et al's (2016) "perfect-balance assumption" (p. 1784) is certainly implausible in the specific context of the poor design illustrated in Fig. 1. However, more generally, invariance is a scientifically plausible, if not essential, concept (Rouder et al., 2009), and hence we do not think this example provides general grounds for rejecting incomplete fixed-effects ANOVA models.

[3] We use Wilkinson and Rogers' (1973) nomenclature for linear models with the intercept implicit and discrete factors dummy coded in some manner, including using the ":" operator to indicate an interaction.

Completeness in Combined Fixed-and-Random-Effects Models

What are the implications of the foregoing discussion for vDAHSW's dilemma about incomplete combined fixed-and-random-effects models? Ruling out such models precludes following Barr et al's. (2013) adjuration to use maximal random-effects models. Although we prefer Matuschek et al's. (2017) approach—using model selection to identify the best supported random-effects model—even in the latter approach, rejecting incomplete models a priori makes it impossible to specifically test a fixed effect whose interaction with a random effect (e.g., participants) is included in the random-effects model. That is, because completeness requires dropping both together, one cannot tell which is responsible for any change in the Bayes factor, something that vDAHSW illustrate to be potentially consequential. As with fixed-effects models, however, we think that this restriction, although necessary in cases where interpretability has been definitively shown to be an issue (i.e., the continuous fixed-effects cases detailed above), is not general. Indeed, we suspect that it is likely not the most common case in practice given that random effects are typically nominal (e.g., participants or items) and so there are no scaling issues.

To be specific, suppose we identify a random participants effect with our gender factor in the example in the last section (but with potentially more levels) and the fixed effect of stimulus is as before. All that is required for an incomplete model with no main effect of stimulus, $Z \sim P + S{:}P$ (where P is the random participants effect), is that the average response over participants for each level of the fixed effect is the same (so there is no main effect of S), while the response of individual participants do differ, and differ over the levels of S. This might occur by having two equal sub-groups differing in the direction of the S effect with the same sensitivity or, more plausibly, in more continuously distributed sensitives to S if they balance out. To state an example of the latter case in terms of a generative model, suppose that, for a given participant, performance on the i^{th} trial with the first type of stimulus ($z_{1,\text{i}}$) and the j^{th} trial with the second type of stimulus ($z_{2,\text{j}}$) equal the sum of four

random draws from normal distributions, one, $N(\mu_D, \sigma_D^2)$, determining the difference in performance between the two levels of the stimulus factor for all trials (d), a second, $N(\mu_p, \sigma_P^2)$, determining their overall performance for all trials (m), and two draws from a third, $N(0,\sigma^2)$ corresponding to independent measurement noise effects for the two types of trials, n_i and n_j. Hence, $z_{1,i} = m + d/2 + n_i$ and $z_{2,j} = m - d/2 + n_j$. A random effect (P) must be included in the ANOVA model when $\sigma_P > 0$ and a fixed-by-random effect interaction ($S{:}P$) when $\sigma_D > 0$, and the model has no main effect of stimulus (and hence is incomplete) if $\mu_D = 0$. The latter assumption is analogous to "perfect-balance" in Rouder et al's. (2016) fixed-effect example. Rejecting a model that is incomplete with respect to mixed fixed and random effects on such grounds is again tantamount to not allowing a point-null hypothesis (i.e., $\mu_D = 0$). If instead one is willing to accept the possibility of testing a point null, specific tests of fixed effects unconfounded by random effects are possible.

Conclusions

Our review of the statistical literature makes it clear that incomplete continuous fixed-effects models are fine to use when the predictors are ratio scaled, but that they should not be used with interval-scaled predictors, or with arbitrarily translated ratio-scale predictors, with the specific exception of the sort of "special-point" cases discussed by Nelder (1998). This conclusion has a transparent basis in arguments based on the simple algebraic properties of linear predictor equations. However, we were unable to find in the literature, or to ourselves generate, a rationale for generalizing these arguments to ANOVA models. We hope that this comment will provide grounds for vDAHSW to fill this gap by providing the explicit mathematical basis for rejecting incomplete ANOVA models that exists for continuous regression models. Further, the separate "perfect-balance" argument against incomplete ANOVA models first put forward by Rouder et al. (2016), and repeated by vDAHSW, seems to us to rely on finding point-null hypotheses corresponding to complete models to be meaningful but point-null hypothesis corresponding to incomplete models to be meaningless. Again, we hope that in replying to this comment, vDAHSW can provide an explicit basis for why one is acceptable, and the other is not. Until then, we know of no general basis for recommending against using incomplete ANOVA models either with respect to fixed-effects only or with respect to mixed fixed-and random effects.

Author Contributions The authors contributed equally.

Funding DM is supported by a Vidi grant (VI.Vidi.191.091) from the Netherlands Organization of Scientific Research (NWO).

Availability of Data and Material Not applicable.

Code availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*, 137–181.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bernhardt, I., & Jung, B. (1979). The interpretation of least squares regression with interaction or polynomial terms. *The Review of Economics and Statistics, 61*, 481–483.

Griepentrog, G. L., Ryan, J. M., & Smith, L. D. (1982). Linear transformations of polynomial regression models. *The American Statistician, 36*, 171–174.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Package 'BayesFactor'. Available at https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf. Accessed 21 Sep 2021

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

Nelder, J. (2000). Functional marginality and response-surface fitting. *Journal of Applied Statistics, 27*, 109–112.

Peixoto, J. (1990). A property of well-formulated polynomial regression models. *The American Statistician, 44*, 26–30.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician, 41*, 311–313.

Prince, M., Brown, S. D., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods, 17*, 78–99.

Raftery, A. E., Painter, I. S., & Volinksy, C. T. (2005). BMA: An R package for Bayesian model averaging. *R News, 5*, 2–8.

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review, 23*, 1779–1786.

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.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*, 304–321.

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*, 225–237.

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *Computational Brain & Behavior*.

van Ravenzwaaij, D., Brown, S. D., Marley, A. J., & Heathcote, A. (2020). Accumulating advantages: A new approach to multialternative forced choice tasks. *Psychological Review, 127*, 186–215.

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

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58–76.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 22*, 392–399.