



# When Fixed and Random Effects Mismatch: Another Case of Inflation of Evidence in Non-Maximal Models

João Veríssimo<sup>1,2</sup>

Accepted: 5 August 2022 / Published online: 15 September 2022  
© The Author(s) 2022

## Abstract

Mixed-effects models that include both fixed and random effects are widely used in the cognitive sciences because they are particularly suited to the analysis of clustered data. However, testing hypotheses about fixed effects in the presence of random effects is far from straightforward and a set of best practices is still lacking. In the target article, van Doorn et al. (*Computational Brain & Behavior*, 2022) examined how Bayesian hypothesis testing with mixed-effects models is impacted by particular model specifications. Here, I extend their work to the more complex case of multiple correlated predictors, such as a predictor of interest and a covariate. I show how non-maximal models can display ‘mismatches’ between fixed and random effects, which occur when a model includes random slopes for the effect of interest, but fails to include them for those predictors that correlate with the effect of interest. Bayesian model comparisons with synthetic data revealed that such mismatches can lead to an underestimation of random variance and to inflated Bayes factors. I provide specific recommendations for resolving mismatches of this type: fitting maximal models, eliminating correlations between predictors, and residualising the random effects. Data and code are publicly available in an OSF repository at <https://osf.io/njaup>.

**Keywords** Linear mixed-effects models · Random effects · Model comparisons · Hypothesis testing · Bayes factors

Linear regression has its origins in the work of Sir Francis Galton, who developed it to understand how the characteristics of offspring could be predicted from those of parents (Azen & Budescu, 2009; Senn, 2011). Galton noticed early on that some characteristics appeared to skip generations, that is, they seemed to be inherited from grandparents or previous ancestors. This led him to propose a “general law of heredity” (Galton, 1897, 1898), according to which multiple generations of ancestors influence the traits of an individual, with each generation contributing less and less as they go backwards in time. Galton’s law contained inaccuracies (Bulmer, 1998) and aligned with his eugenic agenda (Paul, 1995), but it also formalised the rudiments of multiple regression (Azen & Budescu, 2009; Stanton, 2001): the notion that outcomes can have *multiple causes*, which

differ in their importance or *weight*, and with some degree of *overlap* in their effects.

More than a century later, multiple regression has become one of the most important tools in the analytical arsenal of the empirical psychologist. It is routinely used in both experimental and correlational studies, and for both explanatory and predictive goals (Azen & Budescu, 2009; Pedhazur, 1997; Yarkoni & Westfall, 2017). The advantages of multiple regression for psychological research are evident: it allows examining the relative contributions of different variables, whether categorical or continuous, on a variety of data types (Baayen, 2010; Balling, 2008; Coupé, 2018). Crucially, multiple regression takes into account the correlations that exist between predictors, so that each effect is ‘controlled’ (i.e., adjusted) for all of the others (Friedman & Wall, 2005; Vanhove, 2021; Wurm & Fiscaro, 2014; but see Westfall & Yarkoni, 2016).

More recently, mixed-effects (or multilevel) multiple regression models have become widely used in the cognitive sciences (Baayen et al., 2008; Judd et al., 2012; Linck & Cunnings, 2015; Meteyard & Davies, 2020), as they are particularly well-suited to the analysis of clustered data, such

✉ João Veríssimo  
jlverissimo@edu.ulisboa.pt

<sup>1</sup> Center of Linguistics, School of Arts and Humanities,  
University of Lisbon, Alameda da Universidade,  
1600-214 Lisbon, Portugal

<sup>2</sup> Department of Linguistics, University of Potsdam, Potsdam,  
Germany

as when a group of participants responds to a set of items. Mixed-effects models can be applied to a full (unaggregated) dataset and readily accommodate multiple predictors at the participant, item, and trial levels, as well as interactions between them. Further, a defining characteristic of mixed-effects models is that they estimate not only the average effects of predictors—the *fixed* effects—but also the extent to which different ‘clusters’ (e.g., the participants or items in an experiment) vary around those average effects—a type of variation that is captured by the *random* effects. Sir Francis Galton had no mixed-effects models at his disposal, but he acknowledged that particular individuals might depart from his general law, which was concerned only with average results. He extended his law by formalising the particular “prepotencies” of individuals in terms of differences relative to “the mean value from which all deviations are reckoned” (Galton, 1897, p. 402). In mixed-effects models, random effects are estimated in essentially the same way: as deviations around the average (fixed) effects. However, as shown below, this characterisation may fail to apply in models that employ multiple correlated predictors, with consequences for statistical inference.

Fitting a mixed-effects model requires specifying not only a set of fixed effects but also a random-effects structure. The latter can be made more or less complex (Barr et al., 2013; Bates et al., 2018; Matuschek et al., 2017) and this flexibility in model specification can grow quickly when several models have to be compared for the purposes of hypothesis testing. In the target article of this special issue, van Doorn et al. (2022) demonstrate how testing for a fixed effect in the presence of random effects is far from straightforward, since there are several adequate options. Specifically, a model with both fixed and random slopes for an effect of interest can be compared to a model that omits the fixed slope only (the ‘Balanced null’ comparison in the target article; see also Barr et al., 2013) or to a model that includes neither fixed nor random slopes (the ‘Strict null’; see also Rouder et al., 2016). Either comparison may be appropriate, depending on the research questions being asked, since each is essentially testing a different hypothesis.

In this paper, I focus on a specific question raised by van Doorn et al. (2022) regarding Bayesian model comparisons with mixed-effects models: “How to cope with a growing model space as the design becomes more complex?”. Whereas van Doorn et al. (2022) examined the suitability of different ‘null’ models (i.e., those that omit the fixed effect of interest), I focus on the appropriate specification of the ‘alternative’ model (i.e., the model that includes the fixed effect of interest). The “growing model space” that is examined here involves a common situation in psychological research: testing for an effect while controlling for a *covariate*, that is, a different

predictor variable that may correlate with the predictor of interest. Indeed, a compelling aspect of mixed-effects models (and of multiple regression, more generally) is the ability to consider a variety of predictors other than those that are of primary theoretical interest, for example, to rule out potential confounds. This is a desirable property because psychological variables often correlate, and moreover, it can be impractical to equate conditions or groups in all potentially confounding variables (Baayen, 2010; Balota et al., 2004; Brysbaert et al., 2016; Keuleers & Balota, 2015).

Below, I demonstrate that in mixed-effects models with correlated predictors, fixed and random effects can *mismatch*. A mismatch is a lack of alignment in interpretation, such that the very same predictor (in the very same model) can acquire a different interpretation when included in the fixed and random parts of a model. For illustrative purposes, I first show how such mismatches can arise in a very simple mixed-effects model. I then present a series of Bayesian model comparisons with (more complex) models that include a predictor of interest and a covariate. To preview, the results show that mixed-effect models that omit random slopes for covariates contain mismatches, which can lead to overconfident quantifications of evidence for an effect of interest.

## When Fixed and Random Effects Mismatch

Mixed-effects models predict outcomes by estimating not only the average effects of predictors, but also the variance around these effects for different clusters of observations. Specifically, random intercepts and slopes capture the extent to which clusters, like participant or item, differ from the fixed intercepts and slopes. As an example, consider a mixed-effects model with a fixed slope for experimental condition and with random slopes for participants (in addition to fixed and random intercepts). In this model, the fixed slope would reflect the average effect of condition, while the random slopes would be calculated as differences between that fixed slope and the predicted effect for each participant—thus capturing the amount of between-participant variation around the experimental effect of interest (likewise, random intercepts would be estimated as differences from the fixed intercept, thus capturing baseline differences between participants).<sup>1</sup>

This characterisation of random effects as differences from the corresponding fixed effects is made evident by

<sup>1</sup> In frequentist models, random effects are not technically estimates. They are more appropriately called ‘predictions’ or ‘conditional modes’ (Baayen et al., 2008; Bates et al., 2015), since only their (co) variance is estimated.

inspecting the following regression equation, which was inherited from the ‘multilevel modelling’ literature (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 2012) and appears in many introductions to mixed-effects models in psychology (Brauer & Curtin, 2018; DeBruine & Barr, 2021; Hoffman & Rovine, 2007; McNeish & Kelley, 2019; Nezlek, 2008; Singmann & Kellen, 2019), as well as in the target article (Model 6 in van Doorn et al., 2022):

$$y_{ij} = \alpha_i + x_{ij} \cdot \theta_i + \epsilon_{ij}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

In this equation, the outcome  $y_{ij}$ , observed for participant  $i$  in trial  $j$ , is predicted by  $x_{ij}$  (this may denote, for example, the experimental condition that was seen in that particular trial). The intercept  $\alpha_i$  and the slope  $\theta_i$  are both ‘cluster-specific effects’, that is, predicted effects for a given participant  $i$ . These are decomposable into a fixed and a random part:

$$\alpha_i = \mu + u_i$$

$$\theta_i = \nu + v_i$$

The terms  $\mu$  and  $\nu$  are the mean intercept and slope for the whole group (i.e., the fixed effects), and  $u_i$  and  $v_i$  are by-participant deviations (i.e., the random effects), which are assumed to be normally distributed around the fixed intercept and slope:

$$u_i \sim \mathcal{N}(0, \sigma_u^2)$$

$$v_i \sim \mathcal{N}(0, \sigma_v^2)$$

More concretely, consider a model with a fixed effect of condition and with random slopes for each participant. Suppose that the difference between two conditions (i.e., the fixed slope  $\nu$ ) was estimated as 80 ms and the by-participant deviation  $v_3$  (for participant 3) was estimated as  $-12$ ms. Then, the predicted condition effect for participant 3 (i.e.,  $\theta_3$ ) would be 68 ms ( $\nu + v_3$ ).

One important property of this notation is that random effects are explicitly defined in relation to the corresponding fixed effects, as “adjustments” (Baayen et al., 2008), “displacements” (Singmann & Kellen, 2019), or “deviations” (McNeish & Kelley, 2019; Raudenbush & Bryk, 2002) of particular participants *from the fixed-effects means*. While this formulation is indeed accurate in most cases, it obscures an interesting fact about mixed-effects models, namely, that random and fixed effects are defined independently of one another. This can be seen in the alternative matrix-form notation of the mixed-effects equation (e.g., Baayen et al., 2008; Bates et al., 2015; Demidenko, 2013):

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \boldsymbol{\epsilon}_i$$

Here,  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are the fixed-effects ‘model matrix’ and a vector of regression coefficients, respectively, and  $\mathbf{Z}$  and  $\mathbf{b}_i$  are the random-effects model matrix and a vector of deviations for participant  $i$ . As shown in the equation, fixed and random effects have their own independent model matrices, which can consist of different sets of explanatory variables. When that is the case, fixed effects and random deviations may end up acquiring mismatching interpretations.

## An Illustrative Example

To illustrate how mismatches between fixed and random effects can arise, consider the following simulated dataset of 100 hypothetical participants responding to 100 trials in each of two experimental conditions (A and B) in a fully-repeated design. Random effects were generated from a multivariate normal distribution, with a ‘true’ random intercept standard deviation (SD)  $\sigma_u = 1$ , random slope  $\sigma_v = 1$ , and intercept–slope correlation of  $-1$ . These parameters produce the (admittedly unrealistic) random-effect deviations depicted in Fig. 1. As shown in the figure, there is substantial between-participant variation around the mean of condition A, but a complete lack of variation in condition B. Note that the parameters in this simulation were chosen as to clearly reveal a mismatch between fixed and random intercepts, but the discussion in this section does not depend on these particular values.

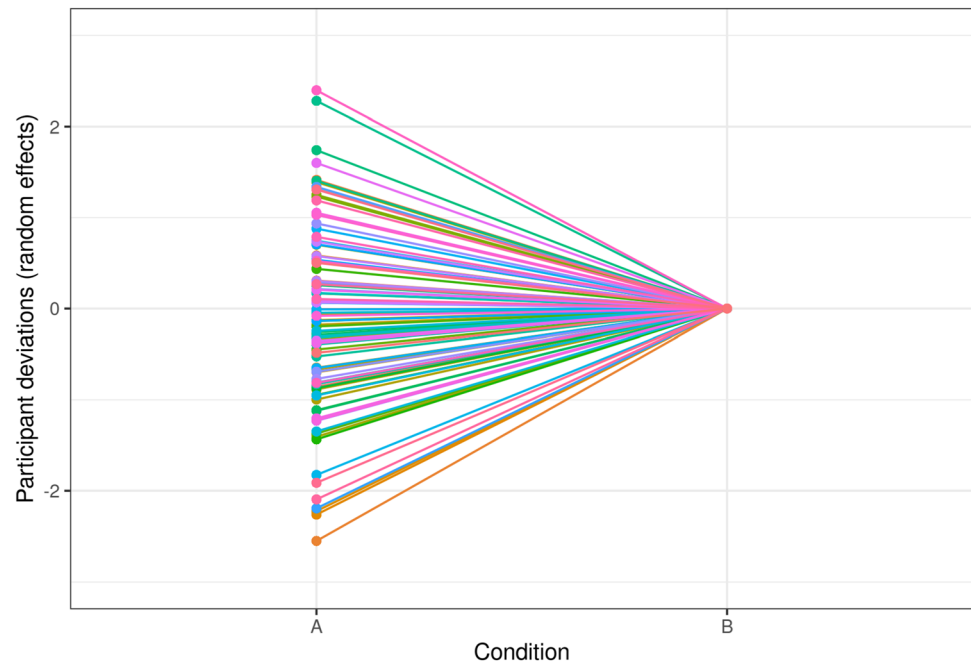
This dataset was first analysed with a ‘maximal’ mixed-effects model with random intercepts and slopes for participants (Barr et al., 2013). In the syntax of the *lme4* R package (Bates et al., 2015):

```
Response ~ 1 + Condition +
(1 + Condition | Participant)
```

Two different versions of this maximal model were fitted. In the first one, condition was dummy-coded as A=0 and B=1.<sup>2</sup> The results showed that in this model the SD of the random intercepts was estimated as (approximately) 1, which was the true random SD in condition A; likewise, the estimated fixed intercept corresponded to the mean of condition A. In the second version, the model was refitted after reversing the coding of condition: condition B was now coded as 0 and condition A as 1. In this version of the model, the random-intercept SD was estimated as (approximately) 0, reflecting the lack of between-participant variation in condition B, and as expected, the fixed intercept reflected the mean of condition B. In sum, in both versions of the model, the interpretations of fixed and random intercepts were

<sup>2</sup> Dummy-coding is also called ‘treatment’ coding. In R, this is the default contrast coding that is applied to factors.

**Fig. 1** Simulated random-effect deviations in two conditions (A, B), with true between-participant SD of 1 for condition A, and 0 for condition B



well-aligned, since both referred to the particular condition that was coded as 0. In such cases, random effects can indeed be seen as deviations from the corresponding fixed effects.

In a second analysis, I fitted a model with random intercepts only, but no random slopes:

```
Response ~ 1 + Condition + (1 | Participant)
```

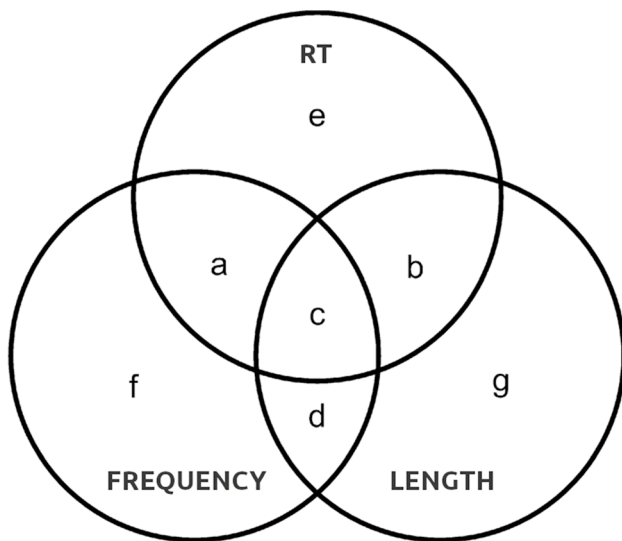
Again, two different versions were fitted, one in which condition was coded as A=0 and B=1, and one with the reverse coding. The results differed from those of the maximal model in two interesting ways. First, the SD of the random intercepts was estimated to be exactly the same in the two versions, irrespective of how condition was coded—this is despite the fixed intercept producing the same estimates as before, which varied with the particular coding that was employed. Second, the SD of the random intercepts was estimated as (approximately) 0.5, which was the average between-participant variation across the two conditions. Thus, in this model, it cannot be said that random effects are deviations from the corresponding fixed effects. If that were the case, the between-participant SD would have been 1 (when the intercept referred to condition A) or 0 (when it referred to condition B), but not 0.5, as was obtained.

This pattern of results can be explained by recognising that the model without random slopes contains a *mismatch* between the interpretations of fixed and random intercepts. The fixed intercept corresponds to the mean outcome when all predictors were at 0, and so it necessarily depended on

the reference level of condition (A or B). Random intercepts, in contrast, were not constrained by the value of condition, because condition was not included as a predictor in the random-effects model matrix. As a result, random intercepts no longer varied around the fixed intercept, but instead, around the *grand mean*! One can visualise this by imagining the midpoint between A and B in Fig. 1; at that point, the between-participant variation corresponds to the average variation in the two conditions (hence, to an SD of 0.5).

The mismatch can also be illustrated by examining the random effects for particular participants. For example, Fig. 1 shows that the participant with the largest outcome in condition A had a ‘true’ random-effect deviation of 2.40 (i.e., their true outcome in condition A was 2.40 above the group mean). In the maximal model, the deviation for this participant was estimated as 2.43—very close to the true value—whereas in the model without random slopes, it was 1.21, a striking underestimation. Again, the reason is that in the model without random slopes, the random intercept for this participant no longer referred to a deviation from the mean of condition A; instead, it was an estimate of how much this participant’s mean outcome (across both conditions) differed from the grand mean.

To sum up, estimates of random effects can differ substantially depending on which other predictors are included in the random-effects model matrix. In the simple example above, mismatches took place at the level of the intercept, such that different interpretations of random intercepts were obtained when random slopes were present or absent. Below, I show that random slopes can also show



**Fig. 2** Venn diagram illustrating the effects of two continuous predictors, frequency and length, on RTs, with both predictors included in a multiple regression model. The sections labelled with small letters (a to f) represent the shared and unique parts of each variable's variance

mismatches, depending on whether *other* random slopes—for covariates—are included or omitted.

### When Do Mismatches Occur?

I propose that mismatches arise when the following conditions are present, *both of which need to be satisfied*: (a) when the model matrices for fixed and random effects contain different sets of predictors, which is the case in ‘non-maximal’ models; and (b) when predictors are not independent, such that the inclusion or omission of one changes the interpretation of the others. The example above meets both conditions. First, there was a mismatch only when fitting a model without random slopes, that is, when the model matrices—the ‘formulas’—for fixed and random effects contained different predictors. Second, the interpretation of the intercept depended on the presence or absence of another predictor and on its particular coding. More specifically, on the fixed-effects side, a dummy-coded predictor constrained the fixed intercept to represent its reference level; on the random-effects side, the absence of slopes led the random intercepts to capture the between-subject variation around the grand mean (not around the reference level), thus acquiring a distinct interpretation from their fixed-effects counterpart.

Whereas the previous example was based on mismatches between fixed and random intercepts, the remainder of this paper will focus on mismatches between fixed and random slopes, and more specifically, on how inferences about a fixed effect of interest can be impacted by the presence or absence of random slopes for a covariate. A more concrete examination of the mechanics of multiple regression will be

useful to understand the nature of mismatches in such situations, as well as their consequences for hypothesis testing. Consider an experiment in which a researcher is interested in testing the effects of word frequency and word length on lexical decision response times (RTs). Figure 2 displays a Venn diagram of the relations between the different variables, in which circles represent their variance (see O’Brien, 2018; Wurm & Fiscaro, 2014; York, 2012; for similar figures). In this example, frequency has an effect on RTs, with more frequent words being responded to faster. The portion of RT variance that is accounted for by frequency is represented by areas *a* and *c* (taken together). Length has an effect on RTs as well, with longer words being associated with slower responses; this effect is represented by areas *b* and *c* together. However, frequency and length also correlate with one another (not only with RTs), since more frequent words tend to be shorter. Their negative correlation is represented by areas *c* and *d* together. The researcher may then wonder whether the effect of frequency on RTs remains (to some extent), even after accounting for the effect of length—or conversely, whether the effect of length is detectable beyond the effect of frequency. As mentioned above, a common solution for disentangling the effects of several correlated predictors is to consider them simultaneously, in a multiple regression model. Indeed, when both frequency and length are included as predictors, the effect of frequency on RTs corresponds to area *a* (only), the effect of length to area *b* (only), and their shared effect (area *c*) is not attributed to either predictor (despite being part of the model’s overall explained variance). In this way, coefficients in a multiple regression model reflect the *unique* contribution of each predictor, over and above all others (Baayen, 2010; Friedman & Wall, 2005; Wurm & Fiscaro, 2014; York, 2012).

In a mixed-effects model the complexity of these relationships increases, because predictors can be included both as fixed and random slopes. For example, the frequency predictor can be included not only as a fixed effect, capturing the average effect of frequency, but also as a by-participant random slope, reflecting the belief that the effect of frequency differs across participants. However, suppose that length is treated as a fixed covariate, but is left out of the random-effects structure, as such:

$$\text{Response} \sim 1 + \text{Frequency} + \text{Length} + (1 + \text{Frequency} \mid \text{Participant})$$

The model matrices for fixed and random effects would consist of different predictors, since the fixed effects comprise frequency and length, but the random slopes include only frequency. As in the example above, in which the interpretation of the intercept was altered by the presence of a dummy-coded predictor, here too the presence of a covariate (length) alters



**Table 1** Fixed and random effects in the different alternative models ( $\mathcal{M}_4$  to  $\mathcal{M}_1$ )

	Fixed effects			Random effects		
	1	2	3	1	2	3
<b><math>\mathcal{M}_4</math>: Maximal</b>						
Predictors	Intercept	V1	V2	Intercept	V1	V2
Interpretation	Grand mean	V1 adjusted for V2	V2 adjusted for V1	Grand mean	V1 adjusted for V2	V2 adjusted for V1
<b><math>\mathcal{M}_3</math>: Reduced</b>						
Predictors	Intercept	V1	V2	Intercept	V1	—
Interpretation	Grand mean	V1 adjusted for V2	V2 adjusted for V1	Grand mean	V1 (unique + shared)	—
<b><math>\mathcal{M}_2</math>: Residualised</b>						
Predictors	Intercept	V1	V2	Intercept	V1 residualised	—
Interpretation	Grand mean	V1 adjusted for V2	V2 adjusted for V1	Grand mean	V1 adjusted for V2	—
<b><math>\mathcal{M}_1</math>: Random-intercepts-only</b>						
Predictors	Intercept	V1	V2	Intercept	—	—
Interpretation	Grand mean	V1 adjusted for V2	V2 adjusted for V1	Grand mean	—	—

the interpretation of the frequency effect, by virtue of their correlation. We can thus distinguish two frequency predictors, which are nominally the same, but have differing interpretations: ‘fixed frequency’ is adjusted for length and corresponds to area *a* in Fig. 2; ‘random frequency’ (or more precisely, the estimate around which the random slopes for frequency vary) is *unadjusted* for length and corresponds to areas *a* and *c* together. Again, there is a mismatch between fixed and random effects.

What are the consequences of such mismatches for statistical inference? Estimates of random variation play a critical role in mixed-effects models, since they determine how much uncertainty there is around the fixed effects. For example, models that lack random slopes for an effect of interest can produce artificially lowered standard errors and an increased rate of Type I errors (Barr et al., 2013). Similarly, in Bayesian hypothesis testing, evidence for a fixed effect can be dramatically inflated when the random effects fail to account for all sources of variation; a clear demonstration is provided in van Doorn et al.’s simulation of a random-intercepts-only model fitted to unaggregated data (Example 1 in van Doorn et al., 2022; see also Oberauer, 2022). I hypothesised that mismatches between fixed and random effects are a type of misspecification that, much like the omission of random slopes for an effect of interest, can lead to an inflation of evidence in hypothesis testing. More specifically, I expected that non-maximal models that include random slopes for an effect of interest (e.g., frequency), but that *fail to include them for covariates* (e.g., length), can underestimate the uncertainty around a fixed effect of interest, and in turn, produce inappropriate Bayes factors in model comparisons with mixed-effects models.

### Model Comparisons With Synthetic Data

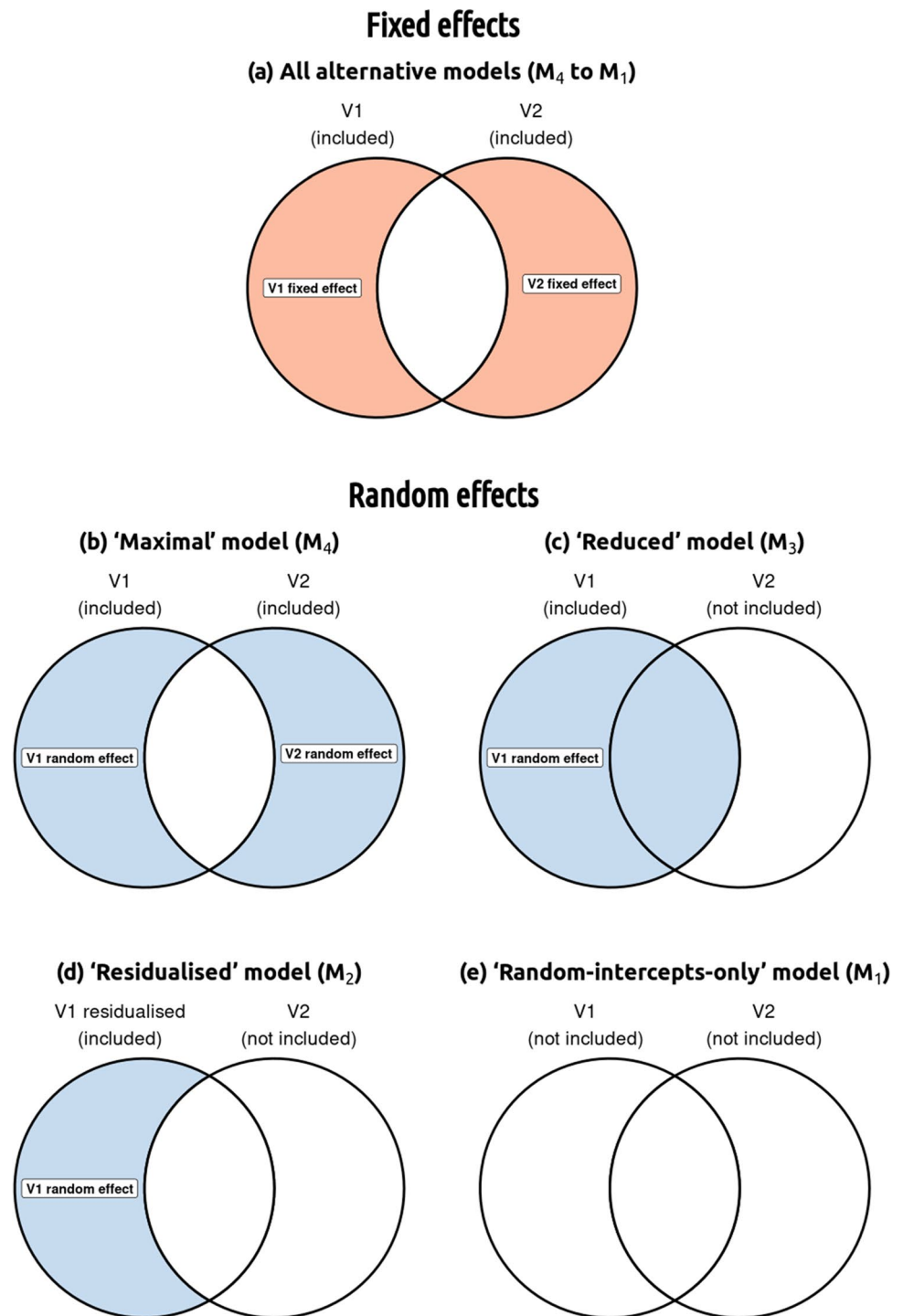
I examined the impact of mismatches on statistical inference through Bayesian model comparisons with a synthetic dataset. The dataset was generated from the effects of two correlated predictors, V1 and V2, analogous to frequency and length in the example above. V1 and V2 were both defined as item-level, within-participant continuous predictors, with V1 as the effect of interest and V2 treated as a covariate. The effects of V1 and V2 varied in magnitude across participants—some participants showed larger V1 or V2 effects and others smaller—a source of variation that can be captured by an appropriately specified random-effects structure. I fitted a set of Bayesian mixed-effects regression models featuring distinct random-effects structures, in order to investigate not only the negative consequences of mismatches, but also the ways in which they can be ameliorated.

### The Candidate Models

Bayesian hypothesis testing is performed by specifying an *alternative model*, which includes the effect of interest, and a *null model*, which does not include the effect, thereby setting the corresponding parameter to zero. Bayes factors quantify the evidence that the data provides in favour of one versus the other model (as the ratio of their marginal likelihoods), and thus allow inferring about the existence or absence of an effect (Mulder & Wagenmakers, 2016; Rouder et al., 2009, 2018; Schad et al., 2021).

I defined a set of *four* alternative models,  $\mathcal{M}_4$  to  $\mathcal{M}_1$  (with  $\mathcal{M}_4$  having the most complex random-effects structure and

**Fig. 3** Venn diagrams representing the unique and shared parts of the fixed effects of V1 and V2 in all four alternative models (panel a), and the random-effects structures in each of the four alternative models (panels b to e), with each shaded area showing the interpretation of the corresponding random slopes (i.e., the effects around which the random slopes for V1 and V2 vary). In these diagrams, the response variable was omitted for reasons of simplicity



$M_1$  the least complex). I first describe all four alternative models, and then the null models that they were compared to. The fixed and random effects included in each of the alternative models are displayed in Table 1. For each model, the first row shows the actual predictors that were included in the (fixed and random) model matrices, and the second row describes their interpretation (i.e., the contribution of

each predictor when taking into account its correlation with the other).

All four models contained the same fixed effects: an intercept, the V1 continuous predictor (which was the fixed effect of interest), and the V2 continuous covariate. Since V1 and V2 were jointly considered in all four models, estimates for V1 reflect the effect of V1 when adjusted for the V2 covariate, or in other words, the unique contribution of V1

over and above that of V2. Figure 3 (panel a) represents the unique and shared effects of V1 and V2 through a Venn diagram (in this figure, the response variable was omitted for reasons of simplicity). Note that none of the four models specified an interaction between V1 and V2 (likewise, the data was generated without such an interaction between predictors; see below). As can be seen in Table 1, the four alternative models differed in their random-effects structures. The random effects included in each model are also represented with Venn diagrams in Fig. 3 (panels b to e), with each shaded area showing the interpretation of the random slopes (i.e., the effects around which the random slopes for V1 and V2 are allowed to vary), for each of the four models.

The first alternative model,  $\mathcal{M}_4$  was the true model from which data was generated (see below for details). This was a ‘Maximal’ model that included all random effects justified by the design (Barr et al., 2013), namely, random intercepts and random slopes for both V1 and V2. In *lmer* syntax:

```
Response~1 + V1 + V2 + (1 + V1 + V2 | Participant)
```

Since the random effects exactly paralleled the fixed effects, model  $\mathcal{M}_4$  did not contain any mismatches. The fixed effect of V1 reflects its effect when adjusted for the V2 covariate and the random slopes for V1 capture the between-participant variation around that same V2-adjusted effect (see Fig. 3, panels a and b). In this way, model  $\mathcal{M}_4$  can appropriately estimate the variation around the effect of interest, and thus serves as an appropriate benchmark against which the other models can be assessed.

Model  $\mathcal{M}_3$  was a ‘Reduced’ model that employed a reduced (i.e., non-maximal) random-effects structure, from which the random slopes for the V2 covariate have been removed (see also Table 1):

```
Response~1 + V1 + V2 + (1 + V1 | Participant)
```

Since V1 and V2 are correlated, accounting for the V2 covariate alters the interpretation of V1. Crucially, in model  $\mathcal{M}_3$  this happens only on the fixed-effects side, because the random slopes for V2 have been left out. Thus, model  $\mathcal{M}_3$  contains a mismatch between fixed and random effects, with the same predictor acquiring distinct interpretations in the fixed and random parts of the model. More specifically, the fixed estimate for V1 reflects its unique contribution, adjusted for V2, but the random slopes for V1 capture the between-participant variation around the *unadjusted* V1 effect (i.e., around an effect that also contains its shared part with V2; see Fig. 3, panel c). I hypothesised that this model would incorrectly estimate the between-participant variance around V1—despite including random slopes

for V1—which could in turn produce an inflated assessment of the evidence in favour of the V1 effect.

The next alternative model,  $\mathcal{M}_2$ , was dubbed the ‘Residualised’ model. As in model  $\mathcal{M}_3$ , the ‘Residualised’ model has also undergone a simplification of its random-effects structure, by removing the random slopes for the V2 covariate. However, the V1 predictor has now been *residualised* (Wurm & Fiscaro, 2014; York, 2012) prior to its inclusion in the random-effects model matrix (see Table 1). The rationale for this procedure is that the mismatch that arises in the ‘Reduced’ model can be eliminated by recoding the V1 random predictor in a way that keeps only its unique part. In order to residualise a predictor, one first fits a regression in which the predictor is itself regressed on the other variables with which it correlates:

$$V1 \sim 1 + V2$$

The residuals of this model are the part of V1 that is not explainable by the V2 covariate. These residuals are then used in place of the original random predictor, as such:

```
Response ~ 1 + V1 + V2 + (1 + ResidualisedV1 | Participant)
```

Such a specification may appear odd because different predictors are included as fixed and random effects. Nevertheless, I expected that this procedure would bring about an alignment between fixed and random effects, since residualising V1 amounts to keeping only its unique part in the random-effects model matrix (see Fig. 3, panel d). In this way, model  $\mathcal{M}_2$  contains a V1 predictor that is adjusted for V2 on both the fixed-effects side (by virtue of both predictors being included in the model matrix), as well as on the random-effects side (through residualisation). Thus, model  $\mathcal{M}_2$  may in fact be more appropriate than model  $\mathcal{M}_3$  at estimating the between-participant variation around the (V2-adjusted) effect of V1.

The final alternative model,  $\mathcal{M}_1$ , was a ‘Random-intercepts-only’ model. It contained random intercepts for participants, but no random slopes (see Table 1):

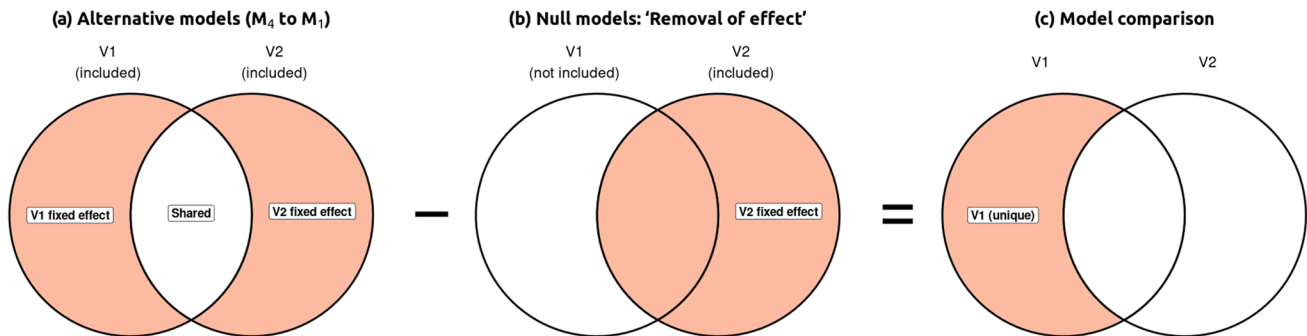
```
Response~1 + V1 + V2 + (1 | Participant)
```

In this model, there were no mismatches between fixed and random effects, given that the only random predictor was the intercept. However, since model  $\mathcal{M}_1$  assumed no between-participant variation around the effect of interest and given that such variation was indeed present, I expected that this model would inflate the evidence in favour of V1 (Oberauer, 2022; van Doorn et al., 2022). The ‘Random-intercepts-only’ model can thus serve as a yardstick for the potential consequences of model misspecification, in



**Table 2** Fixed effects in the different alternative and null models

	Fixed effects		
	1	2	3
Alternative models ( $\mathcal{M}_4$ to $\mathcal{M}_1$ )			
Predictors	Intercept	V1	V2
Interpretation	Grand mean	V1 adjusted for V2	V2 adjusted for V1
Null models (Removal of effect)			
Predictors	Intercept	—	V2
Interpretation	Grand mean	—	V2 (unique + shared)

**Fig. 4** Venn diagrams of the comparisons between alternative and null models, visualised as differences between portions of explained variance. The diagrams show the variance explained by the fixed effects of the alternative models (panel a), the variance explained by

the fixed effects of the 'Removal of effect' null models (panel b), and the result of the model comparisons (panel c). In these diagrams, the response variable was omitted for reasons of simplicity

particular with regard to the underestimation of between-participant variance.

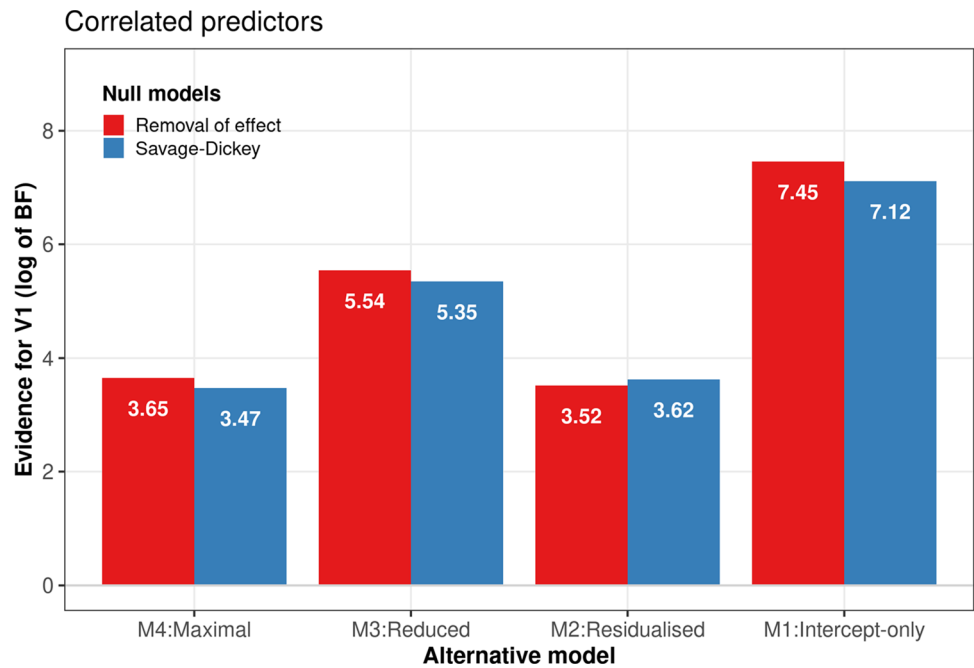
Models  $\mathcal{M}_4$  to  $\mathcal{M}_1$  were compared by means of Bayes factors against four null models, which differed from the alternative models in their fixed effects only (see Table 2). Specifically, each null model was constructed by removing the V1 fixed predictor, so that the only difference between each alternative model and its corresponding null model was the presence or absence of the fixed effect of interest. I will refer to these as the 'Removal of effect' comparisons (in the target article, this is called the 'Balanced null' comparison, which is extended here to the case of correlated predictors, see van Doorn et al., 2022).

One way to understand comparisons between models with correlated predictors is by visualising them as differences between portions of explained variance (Wurm & Fiscaro, 2014). Figure 4 represents the 'Removal of effect' comparisons as a subtraction of Venn diagrams. Note that these diagrams display only the fixed effects, because each null and alternative model had the same random-effects structure (and again, these diagrams omit the response variable for simplicity). As shown in Fig. 4 (and discussed above), the effects of correlated predictors such as V1 and V2 can be decomposed into their unique parts and a shared part. The variance that is explained by their shared part is not

assigned to either predictor, but does contribute to a model's explained variance. Thus, the total variance explained by the fixed effects of alternative models  $\mathcal{M}_4$  to  $\mathcal{M}_1$  is represented by the union of the two circles (see Fig. 4, panel a). The 'Removal of effect' null models remove the V1 fixed effect and leave only the V2 fixed effect (see panel b). The variance explained by the fixed effects of these null models thus corresponds to the whole V2 circle, unadjusted for V1. Finally, the result of the model comparison can be represented as a difference between the two areas, the result of which is the area of V1 that does not overlap with V2 (see panel c). In other words, the model comparison yields the unique contribution of V1, which is the fixed effect of interest.

Besides the 'Removal of effect' comparisons, Bayes factors were also obtained for all alternative models using the Savage–Dickey method (Dickey & Lientz, 1970; Wagenmakers et al., 2010). The Savage–Dickey method allows the computation of Bayes factors from a single alternative model, by calculating the ratio of prior and posterior densities of the parameter of interest at a specific value. When this value is fixed at zero, the Savage–Dickey method is expected to be equivalent to a comparison against a model from which the parameter of interest has been removed, and so I expected this to produce similar results to those of the 'Removal of effect' comparisons.

**Fig. 5** Bayes factors obtained by comparing each of the four alternative models against each of the two null model specifications, for a dataset involving two correlated predictors (see Tables 1 and 2). Bayes factors are logged



**Method**

**Data Generation**

The synthetic dataset consisted of the hypothetical responses of 50 participants  $I$  to 50 items  $J$  with each participant responding to every item. The full dataset of 2500 data points was generated on the basis of the following mixed-effects model equation:

$$y_{ij} = (\mu + u_i) + x_{1j} \cdot (v_1 + v_{1i}) + x_{2j} \cdot (v_2 + v_{2i}) + \epsilon_{ij}$$

The values of the V1 and V2 predictors,  $x_{1j}$  and  $x_{2j}$ , were sampled for each item  $j$  from a multivariate normal distribution with a true correlation between V1 and V2 of  $-0.5$ , and means of 0 and SD of 1 for both predictors. The negative correlation between V1 and V2 reflects the fact that items with a larger V1 value tended to have a smaller V2 value and vice-versa (which is analogous to more frequent words having shorter lengths, for example). Below, I also present an analysis with independent V1 and V2 predictors (i.e., with a true correlation of zero) and show that such a correlation between predictors is a necessary condition for mismatches to arise. The overall mean intercept was  $\mu = 0.5$ . The slope for the V1 effect, which is the effect of interest assessed in the model comparisons presented below, was set at  $v_1 = 0.5$ ; the fixed slope for V2 was  $v_2 = 1$ .

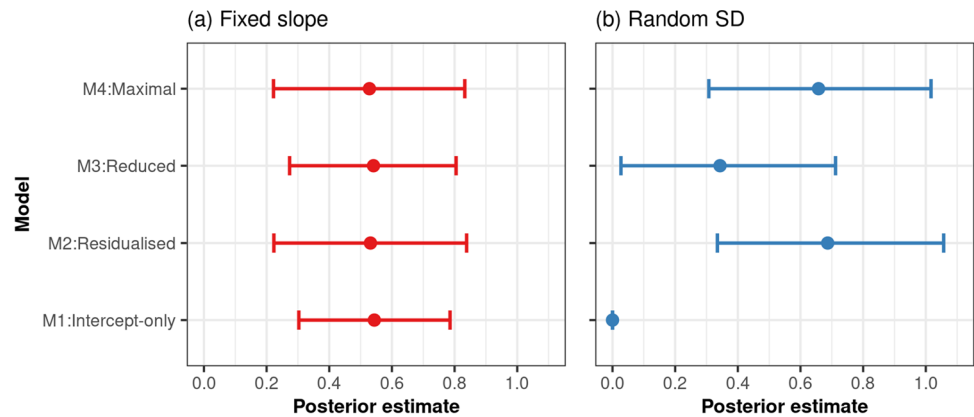
Random intercepts and slopes ( $u_i$ ,  $v_{1i}$ , and  $v_{2i}$ ) were sampled for each participant  $i$  from a multivariate normal distribution with an SD of 0.5 for all random effects (i.e.,  $\sigma_u$ ,  $\sigma_{v_1}$ , and  $\sigma_{v_2}$ ) and correlations between random

effects of .75. The correlation between random intercepts and slopes reflects the fact that participants with a larger mean outcome also tended to show larger experimental effects; the correlation between random slopes indicates that participants with a larger V1 effect also tended to show a larger V2 effect. Such correlations are sometimes obtained in analyses of item-level predictors, since participants may share similar sensitivities to a wide array of item-level variables (e.g., Yap et al., 2012); I return to the issue of the possible role of random-effect correlations in the General Discussion. For simplicity, by-item random effects were not considered in the data generation process. Residual errors  $\epsilon_{ij}$  were sampled for each trial  $i, j$  from a normal distribution with SD  $\sigma_\epsilon = 2.5$ .

**Model Fitting**

A total of eight models were fitted: four alternative models (described in Table 1) and four null models (one for each of the alternative models; see Table 2). Models were fit in a Bayesian framework with the *brms* R-package (Bürkner, 2017). Bayes factors were computed with the `bayes_factor()` function, which calculates the ratio of marginal likelihoods via bridge sampling, or with the `hypothesis()` function, which calculates Bayes factors via the Savage-Dickey method. All priors on fixed effects and variance parameters were Normal (or truncated Normal) distributions with mean of 0 and SD of 1. An *LKJ*(1) prior was used for the correlations between random effects. In order to obtain reasonably stable Bayes factors, all model posteriors contained 60,000 MCMC samples (four

**Fig. 6** Means and 95% credible intervals of posterior distributions for fixed slopes (panel a) and random-effect SDs (panel b) for the effect of interest (V1), in each of the alternative models described in Table 1



chains of 15,000 samples each, besides 2,500 warm-up samples).

## Results and Discussion

The results of the model comparisons are presented in Fig. 5. The figure shows Bayes factors for each of the four alternative models,  $\mathcal{M}_4$  to  $\mathcal{M}_1$  (which differed in their random-effects structures; see Table 1), against two different null model specifications: (a) a model from which the V1 fixed predictor had been removed ('Removal of effect'; see Table 2); and (b) the Savage-Dickey method for the V1 parameter. Following van Doorn et al. (2022), I report the natural logarithms of Bayes factors. A positive value indicates that the alternative model is preferred (providing evidence for the V1 effect) and a negative value indicates that the null model is preferred (providing evidence against the V1 effect).

In all model comparisons, Bayes factors consistently showed evidence in favour of the V1 effect ('very strong' or 'extreme' evidence in the verbal categories of Lee & Wagenmakers, 2013). This was expected since the data was generated from a non-zero V1 effect and the sample size was large enough to yield conclusive Bayes factors. As shown in Fig. 5, Bayes factors produced by the Savage-Dickey method were very similar to those obtained by the comparisons against the 'Removal of effect' null models. Although care should be involved when using the Savage-Dickey ratio (since it may not yield the correct Bayes factor under some circumstances, Heck, 2019; Mulder et al., 2020; Schad et al., 2021), the current results suggest that the ratio of prior and posterior densities at zero is an appropriate way to gauge the Bayes factor for the effect of interest, also in the case of correlated fixed effects.

### Which are the Appropriate Alternative Models?

Although all four alternative models yielded evidence in favour of V1, there were important differences between

them in terms of the *strength* of evidence for the effect (see Fig. 5). Bayes factors for the  $\mathcal{M}_4$  'Maximal' model indicated 'very strong' evidence for V1 (Lee & Wagenmakers, 2013), in both the 'Removal of effect' and Savage-Dickey comparisons. Model  $\mathcal{M}_4$  included all random effects justified by the design (Barr et al., 2013) and, as such, it appropriately captured the between-participant variation around the effect of interest.

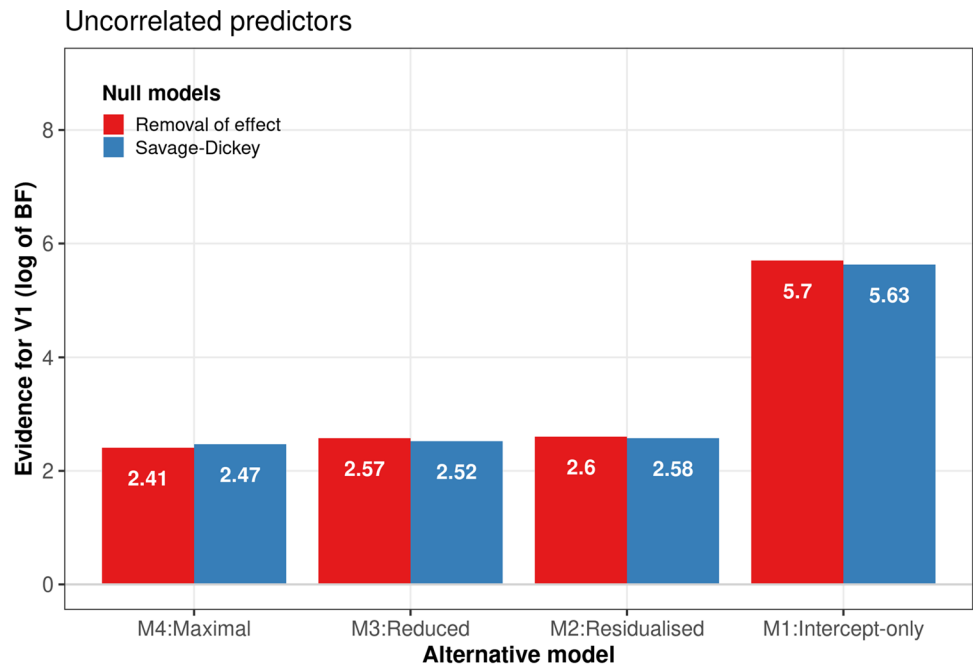
In alternative model  $\mathcal{M}_3$  (the 'Reduced' model), Bayes factors were much larger than for model  $\mathcal{M}_4$ , indicating 'extreme' evidence in favour of V1 (see Fig. 5). Recall that in model  $\mathcal{M}_3$ , random slopes for the V2 covariate were removed, causing a mismatch between fixed and random effects. Specifically, the V1 fixed predictor was adjusted for V2, but the V1 random predictor was not (see Table 1). The larger Bayes factors for this 'Reduced' model than for the 'Maximal' model indicate that the presence of mismatches can inappropriately inflate the evidence for the effect of interest.

In model  $\mathcal{M}_2$  (the 'Residualised' model), V2 was also absent from the random-effects structure, but the V1 random predictor had been residualised to include only its unique part (i.e., the part of V1 that is not explainable by V2). The model comparisons revealed 'very strong' evidence for the effect of interest. Importantly, Bayes factors were lower than those obtained for the 'Reduced' model and very similar to those obtained for the 'Maximal' model (see Fig. 5). Model  $\mathcal{M}_2$  thus yields an appropriate amount of evidence in favour of the effect of interest.

Finally, Bayes factors for model  $\mathcal{M}_1$  (the 'Random-intercepts-only' model) were even larger than for model  $\mathcal{M}_3$ , with extreme support for an effect of V1. Model  $\mathcal{M}_1$  included no random slopes for V1; as expected, this has produced a large inflation of evidence for the effect of interest, relative to a model with a maximal random-effects structure.

Why do models  $\mathcal{M}_3$  and  $\mathcal{M}_1$  inflate the evidence for the effect of interest? An answer can be gathered by examining the posterior distributions of the V1 estimates in the different models. Figure 6 (panel a) shows the means and credible

**Fig. 7** Bayes factors obtained by comparing each of the four alternative models against each of the two null model specifications, for a dataset involving uncorrelated predictors. Bayes factors are logged



intervals (CI) of the posteriors for the fixed effect of V1, in each of the four alternative models. The point estimates were similar across models, but the *width* of the CIs differed. In particular, posterior distributions for  $\mathcal{M}_4$  and  $\mathcal{M}_2$  showed CIs of comparable widths, in model  $\mathcal{M}_3$  the CI was narrower, and in model  $\mathcal{M}_1$  the CI was narrower still. Figure 6 (panel b) shows the estimated random-effect SDs for the V1 random slopes, which reveal that the overconfident estimates in models  $\mathcal{M}_3$  and  $\mathcal{M}_1$  are caused by an underestimation of between-participant variance: in model  $\mathcal{M}_3$  the between-participant variance was estimated to be about one fourth of that in model  $\mathcal{M}_4$  (i.e., the SD was about half), and in model  $\mathcal{M}_1$  it was assumed to be zero (since this model had no random slopes for V1).

In sum, the ‘Random-intercepts-only’ model  $\mathcal{M}_1$  produced inflated Bayes factors for V1 because it assumed no between-participant variance around the effect. This was expected given previous frequentist and Bayesian results (Barr et al., 2013; Oberauer, 2022; van Doorn et al., 2022). But interestingly, a similar failure to appropriately account for between-participant variance was also present in the ‘Reduced’ model,  $\mathcal{M}_3$ . Here, the removal of the V2 covariate from the random-effects structure has led to an underestimation of the random variance around the V1 effect of interest. Note that this underestimation took place despite the random slopes for V1 *still being present*—only the random slopes for the covariate had been removed. However, in the absence of a random slope for the V2 covariate, the interpretation of the random slopes for V1 was altered: in the ‘Reduced’ model, they now captured the between-participant variance around the unadjusted V1 effect, which—given the

particular parameters from which the data was generated—turned out to be much smaller than the variance around the V2-adjusted effect of V1. Ultimately, this resulted in overconfident estimates and severely inflated Bayes factors.

In contrast, the ‘Residualised’ model  $\mathcal{M}_2$  produced appropriate estimates of the between-participant variance around the (V2-adjusted) fixed effect of V1. An unusual property of this model is that it featured different V1 predictors in its fixed- and random-effects parts (see Table 1), since the random-effects structure included random slopes for a *residualised* V1 predictor. Nevertheless, this overt difference between fixed and random predictors is precisely what allowed their interpretations to align and the mismatch to be resolved. This result suggests that residualising a random predictor is an appropriate strategy for eliminating the inflation of evidence that mismatches can produce.

**Uncorrelated Predictors**

If mismatches between fixed and random effects arise only in non-maximal models with correlated predictors, then the inflation of evidence caused by such mismatches may be eliminated when predictors do *not* correlate. This was tested by repeating the data generation, model fitting, and model comparisons presented above, albeit with one single change: the V1 and V2 predictors were now generated from a multivariate normal distribution with a true correlation of zero rather than  $-0.5$ . All other parameters were kept exactly the same in the two analyses.

The results of the model comparisons are presented in Fig. 7. As before, Bayes factors produced by

the Savage-Dickey method were very similar to those obtained by comparing against models from which the V1 predictor had been removed. Also, as was the case with correlated predictors, the ‘Random-intercepts-only’ model  $\mathcal{M}_1$  showed inflated Bayes factors for V1, relatively to the ‘Maximal’ model  $\mathcal{M}_4$ . However, model comparisons involving the ‘Reduced’ model  $\mathcal{M}_3$  now produced Bayes factors comparable to those obtained for model  $\mathcal{M}_4$ . In other words, model  $\mathcal{M}_3$  did not show the inflation of evidence that was previously observed when the two predictors were correlated with one another (cf. Fig. 5; note that Bayes factors were generally smaller for uncorrelated predictors, simply because this was a distinct synthetic dataset).

The different results that were obtained for model  $\mathcal{M}_3$  can be explained by noting that this model only contained a mismatch when V1 and V2 were correlated, but not when they were not. The reason is that when V1 and V2 have a near-zero correlation, then there is virtually no difference between the V2-adjusted effect of V1 (which was included as a fixed effect) and the unadjusted effect of V1 (around which the between-participant random variance was estimated). Likewise, there is virtually no difference between the ‘original’ V1 predictor and the residualised V1, and hence, no difference between the ‘Reduced’  $\mathcal{M}_3$  and the ‘Residualised’  $\mathcal{M}_2$  models. In the case of uncorrelated predictors, both models can accurately estimate the between-participant variation around V1 and, in turn, both yield the appropriate Bayes factor in favour of the effect.

## General Discussion

The present paper has examined Bayes factors for mixed-effects models with multiple correlated predictors. I have demonstrated how non-maximal models that omit random slopes for covariates can display mismatches, with the very same predictor of interest acquiring a different interpretation when it is included as a fixed and as a random effect. This lack of alignment between the fixed and random parts of a model leads to an inaccurate estimation of the amount of random variance around an effect, and ultimately, to an inflated assessment of evidence in its favour.

Previous frequentist and Bayesian results had already shown that models that omit random slopes around an effect of interest yield overconfident estimates, increased rates of Type I errors, and inflated Bayes factors (Arnqvist, 2020; Barr et al., 2013; Heisig & Schaeffer, 2019; Oberauer, 2022; Schielzeth & Forstmeier, 2009; van Doorn et al., 2022). The results presented here provide another demonstration of the negative consequences of non-maximal models. Specifically, models that include random slopes for an effect of interest, but *fail to include them for covariates* (i.e., other predictors

that correlate with the effect of interest), can also display overconfident estimates and an inflation of evidence.

The current work has identified the conditions in which mismatches occur, and in turn, outlined the ways in which they can be resolved. Before making specific recommendations, I discuss whether mixed-effects models with mismatches are common in psychological research and whether the results reported above are generalisable beyond the specific conditions examined in this paper.

## How Common are Mismatches?

The results reported here indicate that mismatches between fixed and random effects occur when a non-maximal model fails to include random slopes for those predictors that correlate with the effect of interest. How common are these two conditions—the omission of random slopes and the presence of correlated predictors—in psychological research?

Even though a well-known recommendation is for researchers to specify models that include all random effects justified by the design (e.g., Barr et al., 2013; Heisig & Schaeffer, 2019), non-maximal models are still routinely employed for a variety of reasons: (a) non-convergence (or singularity) is a familiar problem in mixed-effects models, which can be addressed by the iterative simplification of a model’s random-effects structure (Bates et al., 2018; Brauer & Curtin, 2018; b) maximal models can lead to a loss of statistical power, so researchers may prefer to fit more parsimonious models, with a (simpler) random-effects structure that is “supported by the data” (Matuschek et al., 2017; c) considerations of computational complexity may require fitting models with a smaller number of variance components, especially when conducting fully-Bayesian analyses, which demand more processing time and greater computational power; and (d) researchers may make a distinction between ‘critical’, theory-relevant predictors, which are included as both fixed and random effects, and ‘control’ predictors, which may be left out of the random-effects structure in the interest of parsimony. All of these reasons are in principle justifiable and can lead to models in which random slopes are omitted.

The second condition (i.e., the presence of correlations between predictors) is also prevalent in psychological research. Indeed, many studies in psychology simultaneously examine the contributions of multiple participant- and item-level predictors and these can display substantial correlations between them. For example, at the participant level, demographic variables (e.g., age, years of education, SES), cognitive measures (e.g., working memory, IQ), and language profile variables (e.g., language proficiency, years of exposure) all correlate moderately or even strongly with one another (e.g., Johnson et al., 2013; Veríssimo et al., 2021). If a mixed-effects model includes a set of such



participant-level predictors as fixed effects, but not all such predictors are included as random effects (in this case, as random slopes *by item*), then mismatches may occur. A real example can be found in my own work (Bosch et al., 2019; Verissimo et al., 2018): response latencies in second language processing tasks were predicted by multiple by-participant variables (e.g., age of acquisition, usage, proficiency), but only the theory-relevant predictors were considered for inclusion as by-item random slopes. This distinction between critical predictors (potentially included as random effects) and less important covariates (not even considered for inclusion) may result in models with mismatches, and possibly, in biases in the estimation of random variances.

Similarly, at the item level, the presence of correlations between predictors is a very common situation. A case in point is psycholinguistic experiments examining multiple lexical properties (such as word length, frequency, and neighbourhood), which tend to show strong correlations with one another (Balota et al., 2004; Balling, 2008; Yap et al., 2012). Again, if such predictors are considered as fixed effects but not as random effects, mismatches are bound to occur and may distort the amount of evidence for the effects of interest. Real examples can be found in the studies of Lemhöfer et al., (2008) and Baayen et al., (2007), in which mixed-effects models were used to predict word recognition latencies from various word-level predictors (e.g., frequency, length, entropy), but only some of those predictors were additionally included as by-participant random slopes.

Finally, correlations between predictors are also found whenever confounds in an experimental (factorial) manipulation are accounted for statistically through the inclusion of (continuous) covariates (see Baayen, 2010). In psycholinguistic research in sentence processing, for example, it is common for researchers to compare reading times to different words or regions of text, even if these have not been perfectly equated in their lexical properties. Moreover, due to the incremental nature of sentence processing, differences between conditions can also arise at the trial level, such as when reading times in a region of interest are affected by the processing of the preceding sentence material ('spillover' effects; Mitchell, 1984). To address such differences between conditions, researchers often treat the unmatched lexical properties or the previous-region reading time as continuous covariates (e.g., Avetisyan et al., 2020; Cunnings, 2012; Vasishth, 2006). Crucially, if those covariates are not all included as random slopes in a mixed-effects model, then mismatches may arise. A real example can be found in a self-paced reading study by Shantz (2017), in which only the main experimental manipulations were included in the random-effects structure, but various other fixed covariates (such as length and previous-region reading time) were not.

Given that correlated predictors are pervasive in psychological research and given that there are justifiable reasons

to fit simplified (non-maximal) mixed-effects models, mismatches between fixed and random effects may be the norm, rather than the exception.

## Generality of the Results and Limitations

Although mismatches are likely to be quite common in real-world analyses, it is less clear that they are always consequential, or that they frequently lead to inflations of evidence. The model with mismatches that was described above produced an inflation of evidence because its random slopes were essentially capturing the 'wrong' source of variance. Specifically, this model's random slopes were estimated around an effect that was unadjusted for its correlation with the covariate, whereas the effect of interest *was* adjusted for this correlation. Given the parameters from which the data was generated, the estimated between-participant variance ended up being smaller than the variance around the (adjusted) fixed effect, thereby producing inflated Bayes factors. However, it is also possible (at least in principle) that different underlying parameters could lead the estimated variance around the unadjusted effect to be similar to or even *larger* than the variance around the fixed effect of interest. In the latter case, Bayes factors would be reduced, not inflated, and one could end up concluding (wrongly) in favour of a null effect.

Whether mismatches have negligible or important consequences and whether they bias the evidence in favour of or against a fixed effect may crucially depend on the parameters from which the data was generated. One such parameter is the magnitude and direction of the correlation between the predictor of interest and the covariate. It is clear from the results above that the two predictors need to be correlated for mismatches to occur, because when the true value of this correlation was zero, the 'Reduced model' (which omitted random slopes for the covariate) no longer produced inflated Bayes factors. Another relevant aspect is likely to be the variance-covariance structure of the random effects. That is, not only the correlation between the values of the predictors, but also the correlation between their random slopes may play a role in determining the consequences of mismatches. In all of the analyses above, the correlation between random slopes was assumed to be large and positive (0.75). While such a correlation is not a sufficient condition for mismatches—again, the 'Reduced' model no longer showed an inflation of evidence when the two predictors were uncorrelated, even though the correlation between random slopes was maintained—it may very well be a necessary one.

Unfortunately, one cannot tell in advance if a model with mismatches will substantially distort the evidence and in which direction. In order to find that out, one would have to compare different models (with and without mismatches, as done above), which would require fitting models without

mismatches anyway. Until their consequences are better understood, the best course of action is for researchers to avoid models with mismatches altogether (see below for specific recommendations on how this can be achieved).

An important limitation of the results is that they were obtained from a single set of models, with particular prior distributions, fitted to a single synthetic dataset. In order to address this limitation, I have conducted a number of (informal) sensitivity analyses. Specifically: (a) in multiple model fits and Bayes factor runs, the results were stable and closely replicated (cf. Schad et al., 2021; b) when employing other prior distributions (more or less informative), the same pattern of differences across models was obtained (albeit with larger Bayes factors for more diffuse priors); and (c) repeated analyses with multiple synthetic datasets revealed a similar inflation of evidence in the model with mismatches (although the magnitude of this inflation showed some variation). In sum, the results do not appear to be specific to the particular simulations that were conducted and are likely to hold more generally. Nevertheless, future work should investigate the generality of these results, in particular through simulation studies that systematically vary the magnitudes of (positive and negative) correlations between predictors, as well as the underlying variance-covariance parameters of the random effects.

The results are also general in the sense that they are expected to hold in both Bayesian and frequentist frameworks and when conducting both hypothesis testing and parameter estimation (see Kruschke & Liddell, 2018; Morey et al., 2014; Cumming, 2014, for a discussion). The estimation counterpart of the inflation of evidence in hypothesis testing was already demonstrated above: the model with mismatches produced overconfident estimates (i.e., with credible intervals that were too narrow), both relative to a maximal model and to a model with a residualised random predictor. Furthermore, in additional analyses (not reported above), I have fitted frequentist versions of the candidate models. These showed that the model with mismatches underestimated the SE of the effect of interest and yielded smaller  $p$ -values in likelihood ratio hypothesis tests. This is not surprising, given that the mismatches described here arise from the general properties of mixed-effects multiple regression, which are essentially the same in the different statistical frameworks.

## Recommendations

In order to avoid the negative consequences of mismatches, the first and most straightforward recommendation is to always fit a maximal model, that is, a model with all of the random slopes justified by the design. Again, I emphasise that the reason mismatches occur is not simply a lack of random slopes for the effect being assessed. Rather, the

consequences of mismatches are more insidious, because models with mismatches do indeed contain random slopes for the fixed effect being tested, but omit the random slopes for a covariate predictor. In order to guarantee that mismatches of the type described above do not arise, models should include random slopes for all fixed effects that correlate with an effect of interest.

It should be noted, however, that maximal models may not *always* be immune to mismatches. The particular random slopes that are included in a maximal model are those that can be estimated within each ‘cluster’ (e.g., within each participant). More generally, a random slope is only identifiable if there are multiple different values of the predictor for each level of a random factor. For instance, a between-participant Group variable with two levels—defining two different groups of participants—cannot be included as a by-participant random slope because an effect of Group cannot be estimated within a single individual. In this way, even a maximal model can contain different sets of predictors as fixed and random effects, and if random slopes are not included for the predictors that correlate with an effect of interest, it might be possible (at least in principle) for mismatches and their negative consequences to arise. More research is needed to identify such edge cases. In what concerns mismatches of the type described above (in which the predictor of interest and the covariate can both be estimated as random slopes), it is safe to conclude that they can be eliminated by fitting maximal models.

In cases where it is not possible or practical to fit a maximal model, a second approach for dealing with mismatches is to eliminate all correlations between predictors. Indeed, the results presented above suggest that predictors other than the effect of interest can be left out of the random-effects structure without any negative consequences, just as long as (a) they do not correlate with the effect of interest, and (b) they are not themselves effects of interest. Methods for reducing or eliminating correlations include scaling continuous predictors, linearly combining multiple fixed effects, and employing data reduction techniques (Harrell, 2015; Kowal, 2021; O’Brien, 2007; Tomaschek et al., 2018). In the case of categorical predictors, using sum-coded (e.g.,  $-0.5/0.5$ ) or orthonormal contrasts (Rouder & Morey, 2012) will generally eliminate correlations between predictor variables. Still, there are situations in which these techniques may not be desirable, especially if the hypotheses that are being tested demand the simultaneous consideration of multiple correlated predictors.

If a researcher does end up fitting a model that omits random slopes for those predictors that correlate with the effect of interest, a third approach is available. As demonstrated above, mismatches can be eliminated by residualising the relevant random effects. However, note that as the number of correlated predictors increases, residualisation can quickly

become too complex and should generally be approached with care (Wurm & Fisicaro, 2014; York, 2012). Moreover, when employing residualisation, authors should explicitly define which portions of variance are being assigned to each predictor, both for fixed and for random effects.

More generally, model comparisons involving correlated predictors should be approached with “an abundance of caution” (Rouder & Morey, 2012). Researchers should explicitly take into account how the models being compared differ from one another and which hypotheses are being tested by each comparison, considering not only the fixed effects, but also the random slopes that best capture the variation around the effects of interest.

Finally, in order to prevent mismatches, one must first detect them. To this end, I recommend that authors comprehensively describe all pairwise correlations between the various predictors in a model, as well as one (or more) multicollinearity metrics for regression models (e.g., ‘tolerance’; O’Brien, 2007; Veríssimo & Clahsen, 2014; ‘condition number’; Belsley et al., 1980; Baayen et al., 2007). As shown above, the explicit consideration of how predictors overlap, for example, by making use of Venn diagrams (O’Brien, 2018; Wurm & Fisicaro, 2014) can assist in identifying mismatches, as well as in the specification of alternative and null models for hypothesis testing.

## Conclusion

As observed by van Doorn et al. (2022), mixed model comparisons can be surprisingly intricate. The models examined here were only slightly more complex than the ones in van Doorn et al. (2022), but sufficed to demonstrate another case of inflation of evidence in mixed-effects models. Specifically, the omission of random slopes for predictors that correlate with an effect of interest is an easy-to-miss pitfall that has demonstrable negative consequences for statistical inference.

**Acknowledgements** I thank Shravan Vasishth and the members of his lab, as well as David Kellen, Henrik Singmann, Johnny van Doorn, Frederik Aust, Julia Haaf, Angelika Stefan, and Eric-Jan Wagenmakers for helpful comments.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work has been funded by the Fundação para a Ciência e a Tecnologia (FCT, Foundation for Science and Technology), grant UIDB/00214/2020, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 317633480 (SFB 1287, Project Q).

**Availability of data and material** The datasets generated and analysed during the current study, as well as all fitted statistical models, are available in the Open Science Framework repository at <https://osf.io/njaup>.

**Code availability** This article was composed as a reproducible manuscript using the *papaja* package for the R programming language (Aust & Barth, 2020). The R Markdown file, as well as all R scripts for data generation and analysis are available in the Open Science Framework repository at <https://osf.io/njaup>.

## Declarations

**Ethical approval** Not applicable.

**Consent to participate** Not applicable.

**Consent to publish** Not applicable.

**Conflicts of interest** The author declares no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution*, 35(4), 329–335. <https://doi.org/10/ggkqs5>
- Aust, F., & Barth, M. (2020). *papaja*: Create APA manuscripts with R Markdown (R Package Version 0.1.0.9942).
- Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, 112, 104087. <https://doi.org/10/ghbtpd>
- Azen, R., & Budescu, D. (2009). Applications of multiple regression in psychological research. *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 285–310). SAGE Publications Ltd. <https://doi.org/10.4135/9780857020994.n13>
- Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1), 149–157. <https://doi.org/10/dk585r>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10/fpb5dz>
- Baayen, R. H., Wurm, L. H., & Aycocock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, 2(3), 419–463. <https://doi.org/10/gkzxfq>
- Balling, L. W. (2008). A brief introduction to regression designs and mixed-effects modelling by a recent convert. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Looking at eyes: Eye-tracking studies of reading and translation processing* (pp. 175–192). Samfundslitteratur.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283–316. <https://doi.org/10/dx72cn>

- 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(3), 255–278. <https://doi.org/10/gcm4wc>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. [arXiv:1506.04967](https://arxiv.org/abs/1506.04967)
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10/gcrnkx>
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.
- Bosch, S., Verissimo, J., & Clahsen, H. (2019). Inflectional morphology in bilingual language processing: An age-of-acquisition study. *Language Acquisition*, 26(3), 339–360. <https://doi.org/10/ggffzm>
- Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, 23(3), 389–411. <https://doi.org/10/gd86gx>
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3), 441–458. <https://doi.org/10/ggb87s>
- Bulmer, M. (1998). Galton's law of ancestral heredity. *Heredity*, 81(5), 579–585. <https://doi.org/10/bgzk68>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10/gddxwp>
- Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, 9, 513. <https://doi.org/10/gddf53>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10/5k3>
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3), 369–382. <https://doi.org/10/f35ngk>
- DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592096511. <https://doi.org/10/gjh9p5>
- Demidenko, E. (2013). *Mixed models: Theory and applications with R* (2nd ed.). Wiley.
- Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, 41(1), 214–226. <https://doi.org/10/fc2nps>
- Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2), 127–136. <https://doi.org/10/dhhxqs>
- Galton, F. (1897). The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society of London*, 61(369–377), 401–413. <https://doi.org/10/cw8wsv>
- Galton, F. (1898). A diagram of heredity. *Nature*, 57(1474), 293–293. <https://doi.org/10/cffjkq>
- Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis. *Springer International Publishing*. <https://doi.org/10.1007/978-3-319-19425-7>
- Heck, D. W. (2019). A caveat on the Savage-Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72(2), 316–333. <https://doi.org/10/gk4zsz>
- Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, 35(2), 258–279. <https://doi.org/10/gf9nkc>
- Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, 39(1), 101–117. <https://doi.org/10/bvnt5m>
- Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., Luck, S. J., & Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology*, 27(2), 220–229. <https://doi.org/10/f4sct8>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69. <https://doi.org/10/f33f3t>
- Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, 68(8), 1457–1468. <https://doi.org/10/gkzxsd>
- Kowal, D. R. (2021). *Subset selection for linear mixed models*. <https://doi.org/10.48550/arXiv.2107.12890>
- Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. <https://doi.org/10/gc3gmn>
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. OCLC: ocn861318341
- Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 12–31. <https://doi.org/10/fj5krc>
- Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, 65(S1), 185–207. <https://doi.org/10/f7c46d>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10/gcx746>
- McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, 24(1), 20–35. <https://doi.org/10/gdnbnx>
- Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 104092. <https://doi.org/10/ggjq5>
- Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods of investigating immediate processes in reading. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research*. Erlbaum.
- Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, 25(6), 1289–1290. <https://doi.org/10/gckf4j>
- Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, 72, 1–5. <https://doi.org/10/btqx>



- Mulder, J., Wagenmakers, E.-J., & Marsman, M. (2020). A generalization of the Savage-Dickey density ratio for testing equality and order constrained hypotheses. *The American Statistician*, 1–8. <https://doi.org/10/gk4zxs>
- Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology: Multilevel analyses. *Social and Personality Psychology Compass*, 2(2), 842–860. <https://doi.org/10/dsn35j>
- Oberauer, K. (2022). The Importance of random slopes in mixed models for Bayesian hypothesis testing. *Psychological Science*, 33(4), 648–665. <https://doi.org/10.1177/09567976211046884>
- O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673–690. <https://doi.org/10/bkrhm3>
- O'Brien, R. M. (2018). A consistent and general modified Venn diagram approach that provides insights into regression analysis (F. Zhou, Ed.). *PLoS ONE*, 13(5), e0196740. <https://doi.org/10/gdkkmf>
- Paul, D. B. (1995). *Controlling human heredity: 1865 to the present*. Humanities Press.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed). Harcourt Brace College Publishers.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications
- Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, 23(6), 1779–1786. <https://doi.org/10/f9hfb4>
- Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1), 102–113. <https://doi.org/10/gc9qfx>
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6), 877–903. <https://doi.org/10/ggsfx9>
- 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(2), 225–237. <https://doi.org/10/b3hsdp>
- Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, 26(1), 103–126. <https://doi.org/10/ghbtt6>
- Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). *Workflow techniques for the robust use of Bayes factors*
- Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, 20(2), 416–420. <https://doi.org/10/bcwwvqw>
- Senn, S. (2011). Francis Galton and regression to the mean. *Significance*, 8(3), 124–126. <https://doi.org/10/gf9b83>
- Shantz, K. (2017). Phrase frequency, proficiency and grammaticality interact in non-native processing: Implications for theories of SLA. *Second Language Research*, 33(1), 91–118. <https://doi.org/10/f9k8j9>
- Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (First, pp. 4–31). Routledge. <https://doi.org/10.4324/9780429318405-2>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed). Sage
- Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3), 3. <https://doi.org/10/gd82dx>
- Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71, 249–267. <https://doi.org/10/gg67xj>
- van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2022). Bayes factors for mixed models. *Computational Brain & Behavior*. <https://doi.org/10/gnrnm8>
- Vanhove, J. (2021). Collinearity isn't a disease that needs curing. *Meta-Psychology*, 5. <https://doi.org/10/gnrk22>
- Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In: *Proceedings of the international conference on linguistic evidence*. Tübingen, Germany.
- Veríssimo, J., & Clahsen, H. (2014). Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese. *Journal of Memory and Language*, 76, 61–79. <https://doi.org/10/ggfgmd>
- Veríssimo, J., Heyer, V., Jacob, G., & Clahsen, H. (2018). Selective effects of age of acquisition on morphological priming: Evidence for a sensitive period. *Language Acquisition*, 25(3), 315–326. <https://doi.org/10/ggffzk>
- Veríssimo, J., Verhaeghen, P., Goldman, N., Weinstein, M., & Ullman, M. T. (2021). Evidence that ageing yields improvements as well as declines across attention and executive functions. *Nature Human Behaviour*. <https://doi.org/10/gmh3bj>
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60(3), 158–189. <https://doi.org/10/btbnnf>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think (U. S. Tran, Ed.). *PLoS ONE*, 11(3), e0152719. <https://doi.org/10/f8wpvb>
- Wurm, L. H., & Fiscaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72, 37–48. <https://doi.org/10/gffnfn>
- Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1), 53–79. <https://doi.org/10.1037/a0024177>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10/gcmrmp>
- York, R. (2012). Residualization is not the answer: Rethinking how to address multicollinearity. *Social Science Research*, 41(6), 1379–1386. <https://doi.org/10/gk9zvk>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.