Journal of Memory and Language 125 (2022) 104334

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Contrast coding choices in a decade of mixed models

Laurel Brehm [*], Phillip M. Alday

MPI for Psycholinguistics, Netherlands

ARTICLE INFO

Keywords:
Contrasts
Meta-science
Mixed effect models
Replication crisis

ABSTRACT

Contrast coding in regression models, including mixed-effect models, changes what the terms in the model mean. In particular, it determines whether or not model terms should be interpreted as main effects. This paper highlights how opaque descriptions of contrast coding have affected the field of psycholinguistics. We begin with a reproducible example in R using simulated data to demonstrate how incorrect conclusions can be made from mixed models; this also serves as a primer on contrast coding for statistical novices. We then present an analysis of 3384 papers from the field of psycholinguistics that we coded based upon whether a clear description of contrast coding was present. This analysis demonstrates that the majority of the psycholinguistic literature does not transparently describe contrast coding choices, posing an important challenge to reproducibility and replicability in our field.

Introduction

In 2008, there was a special issue of the *Journal of Memory and Language* dedicated to mixed effect models (MEMs) and other statistical advances, designed for the target audience of cognitive psychologists and psycholinguists. There were two highly influential papers in this issue: Baayen, Davidson, and Bates (2008) and Jaeger (2008). Each of these papers has been cited over a thousand times to date, and these two papers in particular seem to serve as primers on mixed models for many psycholinguists.

Both papers have a similar focus, which is to motivate an ANOVA-using audience to switch analysis methods. In so doing, both papers highlight ways in which MEMs are superior analysis methods for the types of data used in psycholinguistic studies: data with crossed random effects (two sets of repeated measures, such as participants and items) and data that is not necessarily normally distributed, such as binary (binomial) responses. The influence that these papers and the special issue they appear in has had on the field of psycholinguistics cannot be overstated: this special issue initiated a sea change in analysis techniques, such that the dominant analysis tool in the field is no longer ANOVA but MEM.

However, some additional choices do need to be made in MEMs that are not applicable to ANOVAs, meaning that the push to switch analysis methods has created a likely learning curve for statistical novices—even at the software level, MEMs generally require more coding and more analytic choices than ANOVAs. A now substantial literature has developed on some of the unique features of MEMs and the best practices that should be used in psycholinguistics. This includes approaches to random effect selection (see e.g. Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), how to estimate degrees of freedom for p-value calculations (e.g., the infinite degrees of freedom approximation in Baayen et al. (2008), and discussion around the Satterthwaite and Kenward-Roger approximations implemented in the `lmerTest` package in R, Kuznetsova, Brockhoff, & Christensen, 2017), and the best optimizers to use to fit MEMs in R (see e.g. Bates et al., 2015, the lme4 documentation (https://cran.r-project.org/web/packages/lme4/lme4.pdf) and the GLMM FAQ (Bolker, 2021( ).

Mixed models are also now handled in a number of introductory statistical textbooks (e.g., McElreath's *Statistical Rethinking*, Fox's *Applied Regression Analysis and Generalized Linear Models*, and Kretzschmar & Alday (to appear)), in several more advanced textbooks (Pinheiro & Bates, 2000; Zuur, Ieno, Walker, Saveliev, & Smith, 2009; Gelman & Hill, 2006), and in a recent textbook designed specifically for linguists (Winter, 2019). We recommend these resources, in addition to Jaeger (2008) and Baayen et al., 2008), for learning how to use mixed models. Two recent papers are also especially good resources for beginners: Meteyard and Davies (2020) use a meta-analytic approach to showcase the uncertainties that researchers have about using mixed models and present a set of clear reporting guidelines, and Brown (2021) presents a complete MEM tutorial in the R programming language.

Abbreviations: MEM, mixed-effect model.

[*] Corresponding author.

E-mail address: laurel.brehm@gmail.com (L. Brehm).

https://doi.org/10.1016/j.jml.2022.104334
Received 7 July 2021; Received in revised form 13 January 2022; Accepted 19 April 2022
Available online 18 May 2022

Within this ever-growing literature, one topic has received limited attention: how to code fixed effects in MEMs. This is likely because fixed effect coding generalizes from ordinary least-squares regression. However, since MEMs are often used as a replacement for ANOVA, which does not require the same type of coding choices, it is important to address how and why coding choices for fixed effects are made.[1] Most importantly, we note that the default behavior in R (and other statistical software) can mislead novice users who are looking to treat MEMs as a drop-in replacement for ANOVA.

We focus in the current paper on the topic of contrast coding. In MEMs (and all other regression analyses), one needs to make a choice in how to treat categorical predictors. Contrasts are the numeric values assigned to categorical variables in order to enter them into a regression model. There are multiple sensible ways to perform contrast coding, but the choice that is made has implications for the interpretation of effects in a model. Both Baayen et al. (2008) and Jaeger (2008) explicitly stated that they used treatment coding. While this is a common choice in regression models, this contrast coding scheme does not line up with the inferences afforded by ANOVA models (under the most common Type II or Type III sums of squares), and neither paper dedicated much space to the logic behind their choices. This means that MEMs, as used in these two 2008 papers, do not serve as the drop-in replacement for ANOVA that a naive individual may wish for. When combined with the fact that the default in most statistical software is to use treatment coding, the implication is that individuals in our field may be particularly susceptible to incorrect model interpretation.

The question we ask in this paper is whether the psycholinguistic community understands contrast coding, as measured by whether the papers in the citation network of Baayen et al. (2008) and Jaeger (2008) provide sufficient detail to reconstruct their contrast coding choices. To motivate the problem, we begin with a simulated case study on contrast coding in order to highlight the different inferences afforded for model effect terms under two different coding schemes – treatment coding versus sum coding. This is followed by a series of analyses on how contrast coding is described in the psycholinguistic literature published from 2009 to 2018. These analyses highlight that individuals do not, in general, describe their contrast coding choices in sufficient detail to reconstruct their analyses, but that there are some journals and some sub-topics that do better than others in clear description of contrast use, implying a role for individual researchers, journal editors, and reviewers in promoting best practices. More pessimistically, we find that a large proportion of the psycholinguistic literature does not report contrast coding and therefore is uninterpretable in a strict sense. We end with a set of best practice recommendations and a discussion of the consequences that these practices have had on the field.

What's a contrast?

A linear mixed effect model can be expressed mathematically as:

$$y = X\beta + Zu + \epsilon$$

In other words, a response vector y is equal to a vector of fixed effects β times a model matrix X built from numeric values of predictors, plus a vector of random effects u times a matrix Z of indicators for the grouping variables (e.g. which observations belong to a given participant or item), plus a vector of observation-level errors ('noise') ϵ. Or in other, other words: a mixed model is a mathematical description of a set of lines.[2]

Fitting a regression line is conceptually and statistically easy with predictors that are numbers. In this case, the elements entered into the model matrix X are also numbers: the values associated with independent variables. A regression line with only a single continuous fixed effect (only the $X\beta$ and error portions of the equation) is exactly what it looks like in a simple x-y plot: a line that minimizes the vertical distance between values of x and observed values of y for all observations; in a technical sense, the line is the expected values for y for all values of x.

One can also perform a regression analysis— in other words, fit a line— on data that are categorical by selecting some numeric values to apply to the categories. These values are entered in the model matrix X in the equation above. The numeric values chosen to represent comparisons between categorical predictors are contrasts, and these serve the same purpose as the values associated with any numeric variable: to find the line that minimizes error between values of x and observed values of y. Contrasts allow comparisons to be made between one or more levels of a variable – comparing levels to each other, to the mean value of the variable, or to various combinations of other variable levels. The number of comparisons that can be made for a variable depends on how many levels there are: For any categorical variable with N levels, N-1 contrasts are used in a contrast matrix. This is because there are only N-1 ways of creating independent (orthogonal) comparisons between the groups.

For a model containing a single two-level variable, there are two straightforward contrast coding choices to make up the single contrast vector in the contrast matrix. One choice is to use *treatment* (or *dummy*) coding: setting one level as the *reference level* for the model by assigning it zero and setting one level as the *treatment level* by assigning it one to make the contrast vector (0,1). The other is to use *sum* (or *effect*) coding: setting one level as negative and one positive, with zero as the mean of the two levels, to make the contrast vector (-1,1).[3]

Contrast coding changes what the model intercept reflects, since the model intercept is the y value when all predictors are zero. In sum coding, the intercept is the y value associated with the grand mean of the two cells (the average of the two cells), whereas in treatment coding, the intercept is the y value associated with the reference (zero) level. The effect term in a one-predictor model is then interpreted accordingly. For a model containing a single two-level variable, the predictor term when (-1, 1) sum coding is used will reflect half the change in the y value between the two levels, whereas in treatment coding, it will reflect the increase in the y value associated with the treatment level.

The implications of contrast coding become more striking in more complex models. For a three-level factor A, treatment coding creates two contrast vectors: if the first level is the reference, these would be (0,1,0) and (0,0,1); the model intercept reflects the y value at the reference level. This means that any interactions between A and any other factors also need to be evaluated at the reference level of factor A. Sum coding for factors with more than two levels also requires setting a reference level: if the first level is the reference, the two contrasts would be (-1,1,0) and (-1,0,1). The intercept in this model reflects the grand mean of the three factor levels, and the comparisons reflect the difference between each (non-reference) level and the grand mean. A worked example using sum coding in a more complex model appears in the metascientific study below.

Note that more complex comparisons also become available with factors that have three levels or more, including *Helmert* and *difference* (or *repeated*) contrast coding for ordered factors. Selecting among these various options strategically can eliminate the need to perform post hoc

[1] There is an analog of contrast coding in ANOVA: the type of sums of squares. In many ways, this presents a parallel problem: model results are not interpretable without this information, yet they are often unreported, and the defaults in much statistical software, e.g. Type I SS in base R, are often not desirable for psycholinguistic analyses. The different types of sums of squares can also be expressed as different contrast coding schemes for the regression model underlying ANOVA.

[2] This also holds for a generalized linear mixed model, but the lines are transformed via a link function before the observation-level variability is considered, and the observation-level variability may not be Gaussian.

[3] (-.5,.5) is another variant of sum coding. Results are interpreted in the same way: all that differs is the magnitude of beta values, which are twice as large.

tests, as nicely highlighted by Schad, Vasishth, Hohenstein, & Kliegl (2020). Helmert coding tests whether the differences between increasing or decreasing ordered factor levels are uniform. For example, to examine the interference in picture naming that comes from simultaneously listening to Dutch speech, Chinese speech, or noise, He, Meyer, and Brehm (2021) used Helmert coding to compare the average of the two (more challenging) language conditions to the (easier) noise condition and to then compare the two language conditions to each other. Difference coding is also used for ordered factors, but instead isolates the most theoretically useful pairwise comparisons (the adjacent ones). For example, Breen (2018) used backwards difference coding to test how word durations change when reading the children's book *The Cat in the Hat* aloud based upon decreases in the metric hierarchy level (the combination of syllable stress and word position).

Less intuitively, when a model contains multiple variables, contrast coding for one variable also changes the interpretation of other variables. This is the problem we want to highlight in this paper. Because contrast coding changes the interpretation of the intercept, it therefore also changes the interpretation of all main effects, and all interactions except the highest-order one. This is because effect terms in a model are evaluated when the intercept is equal to zero– so a contrast coding scheme where zero is set to reflect one particular factor level will have a radically different interpretation than one in which zero reflects a combination of several factor levels.

In a sum-coded model (-1, 1 coding), the fact that zero is the average of the two levels means that the effect of factor A is evaluated at the average of the factor B levels. This means that in a sum-coded model, the effect of each factor is coded to reflect how it influences the DV while collapsing across any other factors. This is easier to understand in an example. If, in a model of the time it takes to eat a meal, factor A is *Utensils* and factor B is *Foods*, then the model effect terms will describe the effect of each factor, averaging across both levels of the other. These effects will correspond to the main effects in an ANOVA model: the influence of *Utensils* on eating time, regardless of which food was eaten, and the influence of *Foods* on eating time, regardless of which utensil was used. This type of hypothesis testing is often desired: Main effects are often what psycholinguists wish to evaluate statistically.

In comparison, in a treatment-coded model where one level is set to zero (1, 0 coding), the effect of factor A is evaluated at the reference level (zero level) of factor B. This means that the model effect terms are not main effects, but simple effects. Setting the reference level of factor A *Utensils* to `Fork` and the reference level of factor B *Foods* to `Salad` means that the model intercept will be evaluated at the combination of (Fork + Salad), the effect of *Utensils* on eating time will be evaluated when salad was eaten, and the effect of *Foods* on eating time will be evaluated when a fork was used. Importantly, for many research designs, simple effects are not equivalent to main effects. This means that one must know the contrast coding scheme used in order to interpret a regression model.

Case study: Why contrast coding matters

It's easier to see the impact of contrast coding schemes in a fully-worked example. The R code below generates some simulated data where there is a crossover interaction of factors A (*Utensils*) and B (*Foods*) on a dependent measure *RT*, (speed of eating, in minutes). This data pattern has no reliable main effects (there is no overall effect of utensil choice, averaged across levels of foods) but does have reliable simple effects (there is an overall effect of utensil choice on the speed of eating when looking at one food at a time, and an overall effect of food choice on the speed of eating when looking at one utensil at a time).

We begin the example by loading four packages: `lme4` (Bates et al., 2015) is the package for mixed models, `car` (Fox and Weisberg, 2019) is a package for setting nicely-labeled contrasts (among other things), `jtools` (Long, 2020) and `kableExtra` (Zhu, 2021) provide nicely readable model outputs. We also set a random seed so that results will replicate when the code is re-run.

```r
library(lme4); library(car); library(jtools); library(kableExtra)

set.seed(1012)
```

Next, we set some function inputs. We will draw random values from distributions centered around condition means `SpoonSoup`, `ForkSoup`, `SpoonSalad`, and `ForkSalad`, representing the four combinations of two two-level factors `Utensils` and `Foods`. These will all have the same standard deviation `Groupsd`, corresponding to the usual homoskedacity assumption (the same variance in all conditions). We also define that we want the code to generate 20 participants (`ps`, the eaters in our experiment) and 10 items (`ii`, the different main ingredients in each soup and salad–such as potatoes, beets, pasta, etc.).

```r
SpoonSoup <- 5

ForkSoup <- 10

SpoonSalad <- 10

ForkSalad <- 5



Groupsd <- 2



ps <- 20

ii <- 10
```

We build the structure of a data frame containing *ps* by *ii* observations in each cell of `Utensils` and `Foods` by repeating elements the correct number of times and binding them together as columns in a data frame. Participants are numbers proceeded by `p` and items are numbers proceeded by `i`.[4]

[4] This will force these to be coded as characters in the data frame and then as factors when used as random intercepts, which is the desired outcome.

```
ds <- as.data.frame(cbind(rep(c(rep("Spoon",ps*ii),rep("Fork",ps*ii)),2),
                          c(rep("Soup",ps*ii*2),rep("Salad",ps*ii*2)),
                          paste0("p",rep(1:ps,ii*4)),
                          paste0("i",rep(1:ii,each=ps,times=4))))
colnames(ds) <- c("Utensils", "Foods", "Participant", "Item")
```

Then, we create some simulated data by drawing from a random normal distribution `ps` by `ii` times for each cell of the design.

```
ds$RT <- c(rnorm(ps*ii,SpoonSoup,Groupsd),rnorm(ps*ii,SpoonSalad,Groupsd),
           rnorm(ps*ii,ForkSoup,Groupsd),rnorm(ps*ii,ForkSalad,Groupsd))
```

We also create some random effects – random variance attributed to each participant (each eater) and each item in the study (each main ingredient, here used in both soup and salad), centered around zero, of a magnitude that is a fraction of the overall variance. The overall DV is then composed by adding the original random draw with the random effects per participant and per item for a given observation.

```
psre <- rnorm(ps,0,Groupsd/5)
iire <- rnorm(ii,0,Groupsd/5)
ds$RT <- ds$RT + psre[as.numeric(as.factor(ds$Participant))] +
  iire[as.numeric(as.factor(ds$Item))]
```

Finally, we make sure that R is appropriately treating our variables as factors.

```
ds$Utensils <- as.factor(ds$Utensils)
ds$Foods <- as.factor(ds$Foods)
```

Next, we run two linear mixed effect regression models using the function `lmer()` from the R package `lme4` with the predictors of Utensils and Foods, and random intercepts for participants and items. (In real data, one also needs to consider whether random slopes are justified– we set this issue aside for the current paper. See e.g. Barr et al., 2013; Matuschek et al., 2017).

In the first model, we do not set any contrasts, but we do use the base R function contrasts() to look up what they are. The default coding scheme in R is to use treatment coding with the first level alphabetically as the reference. In this model, there appear to be main effects of `Utensils` and `Foods` such that spoons and soup lead to slower eating overall– but note that these are actually simple effects, because the intercept is set to reflect the zero-level for both variables (the `Fork` + `Salad` cell of the design). The correct interpretation of this model is that there is an effect of `Utensils` when eating salad (spoons are a slower way to eat it), and an effect of `Foods` when using a fork (soup is slower to eat with it). There is also an interaction between them, such that it is slower to eat salad with a spoon and soup with a fork. The model is summarized using the `jtools()` function `summ()` which creates formatted model tables; we suppress the R^2 and p values because defining these requires additional assumptions for linear mixed models.

```
contrasts(ds$Utensils)

##       Spoon
## Fork      0
## Spoon     1

contrasts(ds$Foods)

##       Soup
## Salad    0
## Soup     1

m1 <- lmer(RT ~ Utensils*Foods + (1|Participant) + (1|Item), data=ds)
summ(m1, model.info=F, model.fit=F, pvals=F)
```

	Fixed Effects		
	Est.	S.E.	t val.
(Intercept)	4.81	0.21	23.45
UtensilsSpoon	5.06	0.20	25.19
FoodsSoup	5.13	0.20	25.52
UtensilsSpoon:FoodsSoup	−10.10	0.28	−35.52

	Random Effects	
Group	Parameter	Std. Dev.
Participant	(Intercept)	0.46
Item	(Intercept)	0.33
Residual		2.01

	Grouping Variables	
Group	# groups	ICC
Participant	20	0.05
Item	10	0.03

In the second model, we set sum contrasts. The function `contr.Sum()` from the car package is used to do this because as it provides a useful label set. Here, the label `[S.Fork]` reminds us that we are using sum coding with the `Fork` level as the positive value. In this model, the 'main effects' disappear– because in the first model, what looked like main effects were actually simple effects. On average, it takes the same amount of time in this simulation to eat with a spoon as a fork, and the same amount of time to eat soup as salad. However, the interaction is still present, corresponding with the fact that it is slower to eat salad with a spoon and soup with a fork. Importantly, the random effect terms are also identical in both models. That is because contrast coding does not change the random effects (so long as both models converge), nor does it change the highest-order interaction: only the intercept and other lower-order fixed effect terms.

```
contrasts(ds$Utensils) <- contr.Sum(levels(ds$Utensils))

contrasts(ds$Foods) <- contr.Sum(levels(ds$Foods))

contrasts(ds$Utensils)
```

```
##       [S.Fork]

## Fork         1

## Spoon       -1
```

```
contrasts(ds$Foods)
```

```
##       [S.Salad]

## Salad          1

## Soup          -1
```

```
m2 <- lmer(RT ~ Utensils*Foods + (1|Participant) + (1|Item), data=ds)


summ(m2, model.info=F, model.fit=F, pvals=F)
```

Fixed Effects			
	Est.	S.E.	t val.
(Intercept)	7.38	0.16	44.97
Utensils[S.Fork]	−0.01	0.07	−0.09
Foods[S.Salad]	-0.04	0.07	-0.57
Utensils[S.Fork]:Foods[S.Salad]	−2.52	0.07	−35.52

Random Effects		
Group	Parameter	Std. Dev.
Participant	(Intercept)	0.46
Item	(Intercept)	0.33
Residual		2.01

Grouping Variables		
Group	# groups	ICC
Participant	20	0.05
Item	10	0.03

This pair of models highlights the general problem: running a model without knowing the contrast coding leads to results that it is impossible to draw inferences from. Most problematically, what appear to be main effects can be interpreted by a naive reader or experimenter as simple effects, and vice versa. This is especially the case in analyses that rely on significance testing (instead of a model-fitting approach) when no post hoc testing (i.e., with the `lsmeans` or `emmeans` R packages) is done.[5] In comparison, when the contrasts are clearly described– no matter what they are– then the correct inferences can always be drawn by the reader about the model, and post hoc testing typically is no longer necessary.

Metascientific study

Taking a metascientific approach, we next examined the use of contrasts in the citation network of the two influential 2008 papers (Baayen et al., 2008 & Jaeger, 2008J). This allowed us to compile a set of literature with a psycholinguistic focus using mixed models. Within this sample, we coded whether the authors provided details on their contrast coding choices. We asked how patterns changed over time, whether journals differed, and whether certain sub-fields, as indexed by keyword, have more success than others in correct contrast use.

Method

The first step was to compile a database of papers that used mixed models. We performed a search in the Web of Science database on May 05, 2019 for all papers published in the years 2009 to 2018 that cited either Baayen et al. (2008; N = 2294), only Jaeger (2008; N = 803), or both (N = 520). For each paper, one of the authors or a research assistant assessed (i) whether the paper was accessible by our library, (ii) whether it was in English, and (iii) whether it contained any mixed model analyses. At this stage, 233 papers were excluded (14 were inaccessible, 15 were not in English, and 204 did not contain mixed models), leaving 3384 papers.

Next, the papers were all coded by the first author, or coded by a research assistant and then checked by one of the two authors. The first step was to code whether categorical variables were present (N = 3125 yes, 259 no), These 3125 papers are those for which contrast coding is relevant, and make up the data reported in the rest of the paper. In these papers, we then coded if contrasts were explicitly described for one or more variables (N = 1069 yes, 2056 no) by skimming the methods and results section and performing word searches for the following terms: *contrast, code, level, reference, treatment, dummy*.

We counted contrast descriptions as present if a coding scheme was named (*"We used deviation/Helmert/sum coding"*), if a reference level was marked in the text or in a table (*"The reference level of factor A was Y"*), or if numeric values were mentioned (*"We contrast coded all factors as 0.5, −0.5"*). Contrast descriptions were coded as present even when the coding scheme was nonsensical, if it met the guidelines above (e.g., polynomial contrasts for a variable with two levels).

Contrast descriptions were not coded as present if the authors simply said that they "performed contrast coding" or "centered variables" without any further details, as this does not allow reconstruction of the analysis. The first statement is problematic because it does not describe the contrast coding procedure in sufficient detail to reconstruct the analysis. The term 'contrast coding' is sometimes used as a shorthand notation for 'sum coding' (as opposed to leaving the default treatment contrasts), but note that multiple different contrast coding schemes are always available for variables with more than two levels. Statements like this are therefore needlessly confusing, even when they are correctly used to describe that a two-level variable was sum coded. Especially to a naive user, we believe this to be too opaque to be useful. The second statement is problematic because it is unclear which variables it applies to. Continuous variables are centered by subtracting the mean from each value; this sets the intercept (zero) level to the average value of the variable. In terms of category levels, it is typically not clear what the "center" would be (e.g. what is the center of the variable *common pets* with levels *cat, dog, goldfish*?): centering in this sense is fairly nonsensical. In terms of contrast values, the term centering is sometimes used to mean that a weighted contrast coding scheme was used. In this case, the resulting comparisons are *data* dependent instead of *design* dependent. While there are a few cases where it makes sense to adjust the contrasts for the data (i.e. for certain types of unbalanced data that are missing not at random), weighted contrast coding should be done intentionally and transparently (see Sweeney & Ulveling, 1972; Nieuwenhuis, te Grotenhuis, & Pelzer, 2017).

All data and the code to perform the following analyses appear on https://osf.io/jkpxt/. Note that the data are de-identified in order to protect author identities.

Results

Less than a third of papers describe contrasts clearly

Of the 3125 papers in our data set which used one or more categorical variables, and therefore needed to make a choice about contrast coding, only 1069 described their choice explicitly. In other words: only 34% of papers in a large sample of psycholinguistic literature were fully explicit about which choices were made in their data analysis. The *overwhelmimg majority* of papers either did not describe their contrasts at all, or did so insufficiently. This suggests the potential for an enormous replicability problem: readers cannot tell what choices were made about data analysis, nor whether all conclusions drawn about the data were correct.

In a very strict sense, the lack of clear contrast coding choices means that the the statistics in these 2056 papers– 66% of the psycholinguistic

[5] As discussed elsewhere: we endorse a priori sensible contrast coding over post hoc testing. Post-hoc testing *is* a solution to the problem of incorrect model inference, but comes at the cost of introducing multiple comparisons. Moreover, omnibus post hoc testing is often symptomatic of exploratory research, which needs to be interpreted and reported fundamentally differently from confirmatory research (cf. Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Nearly all sets of a priori hypotheses can be tested using N-1 comparisons when these are chosen carefully.

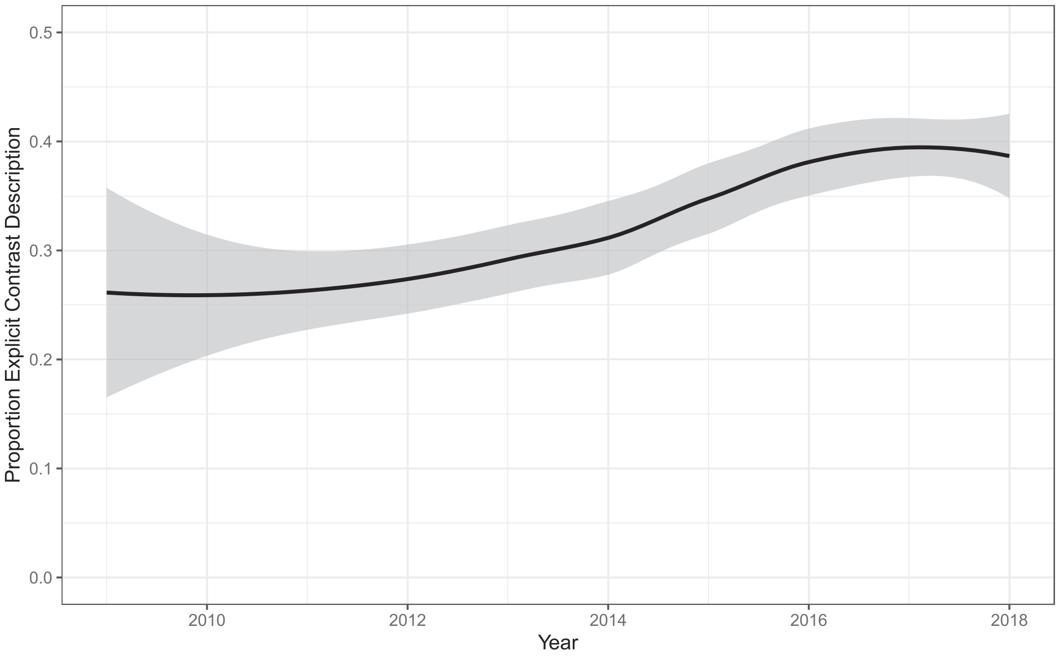

Fig. 1. Proportion of explicit contrast use by year, with loess smooth.

literature sampled– cannot be interpreted. Without knowing the contrast scheme, it is not possible to interpret the model coefficients and associated significance tests – even without interaction terms – because the contrasts are what encodes the hypothesis under consideration. In a loose sense, there is reason for optimism: treatment coding and sum coding, for example, only give different results when interactions are present, and the results are most strikingly different in the presence of crossover interactions, as outlined in the case study above. The larger issue is that we do not know which of these papers make valid conclusions: the statistics presented in the majority of the psycholinguistic literature are, strictly speaking, uninterpretable, because contrast coding choices are not sufficiently described.

Patterns over time are improving

We did observe a general increasing trend over time, as shown in Fig. 1. The contrast description rate has increased to 38.6% by 2018, with the maximum year being 2017 (with 39.4% of papers explicitly describing their contrasts). The implication is that authors may be changing their behavior for the better over time.

We tested this pattern with a generalized linear model, coding year as an ordered factor with orthogonal polynomial contrasts. We selected this contrast coding scheme because it allowed us to test whether trends over time are best described as linear (steadily increasing or decreasing), quadratic (increasing, then decreasing, or vice versa), or some other more complex non-linear pattern (polynomials of third degree and higher). This model appears in Table 1. In this analysis, there was only a significant positive linear trend, such that we would expect future papers from an analogous sample (i.e., those that cite Baayen et al., 2008 or Jaeger, 2008, and which use categorical variables) to be ever more precise with their description of contrast coding.

Table 1
By-year contrast description analysis. Model intercept reflects the grand mean of contrast description (mean of mean descriptions per year), and effects reflect polynomial patterns between year and contrast description, from linear to a 9th degree polynomial. Model formula is: glm(ContrastsUse ∼ Year, family=‘binomial’)

	Est.	S.E.	z val.	p
(Intercept)	−0.78	0.05	−14.68	0.00
Year.L	0.74	0.21	3.57	0.00
Year.Q	0.10	0.20	0.51	0.61
Year.C	−0.18	0.19	−0.95	0.34
Year^4	−0.07	0.18	−0.40	0.69
Year^5	−0.00	0.16	−0.03	0.98
Year^6	−0.04	0.15	−0.27	0.79
Year^7	0.12	0.14	0.86	0.39
Year^8	−0.06	0.13	−0.42	0.67
Year^9	0.09	0.12	0.71	0.48

Standard errors: MLE.

Patterns by journal are varied

To look at how choices about contrast coding might be influenced by journal editors and reviewers, we looked at patterns by journal. There were 567 journals included in the data set, and we extracted the 34 for which we had at least 20 observations. These are shown in Table 2, alongside the abbreviations used in our tables and figures. We selected 20 as our cutoff simply because it is a standard ‘large-enough’ number for many statistical purposes; a higher threshold would have made the sample less reflective of the field as a whole, and a lower one might risk issues with model convergence or overfitting.

This subset of journals was entered into a generalized linear mixed model in which we predicted explicit contrast description by journal with a random intercept for year. Because we treat year in all analyses as a categorical variable, and because it has relatively few levels, it is a valid grouping variable for random intercepts: see e.g. Onkelinx (2017). Within this model, the predictor journal was sum coded, with the median level of journal, when ordered by contrast description rates, set as the (omitted) reference level for the model (this was *Cognitive Science*).

In sum coding with more than two levels, the intercept reflects the grand mean (the mean of all levels of the variable), and each effect reflects whether the level is reliably different from the grand mean; one level must be omitted as a reference level. Setting the reference level as something close to the grand mean means that this omitted comparison is one of the least important ones, and so little relevant information is lost. Using this contrast coding scheme therefore allows us to test whether each journal performs differently than the average of all journals. Model output can be found in Table 3; each contrast in the model is

Table 2
Journals appearing in by-journal analysis, with abbreviations and counts of observation.

Journal Full Name	Journal Abbreviation	Observations
ACTA PSYCHOLOGICA	Acta Psychol	49
APPLIED PSYCHOLINGUISTICS	Appl Psycholinguist	30
ATTENTION PERCEPTION & PSYCHOPHYSICS	AP&P	35
BILINGUALISM-LANGUAGE AND COGNITION	B:L&C	44
BRAIN AND LANGUAGE	Brain Lang	22
COGNITION	Cognit	105
COGNITIVE PSYCHOLOGY	Cog Psychol	20
COGNITIVE SCIENCE	Cog Sci	55
FRONTIERS IN HUMAN NEUROSCIENCE	Front Hum Neurosci	28
FRONTIERS IN PSYCHOLOGY	Front Psychol	170
JOURNAL OF CHILD LANGUAGE	J Child Lang	28
JOURNAL OF COGNITIVE PSYCHOLOGY	J Cog Psych	26
JOURNAL OF EXPERIMENTAL CHILD PSYCHOLOGY	J Exp Child Psychol	24
JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL	JEP:G	29
JOURNAL OF EXPERIMENTAL PSYCHOLOGY-HUMAN PERCEPTION AND PERFORMANCE	JEP:HPP	48
JOURNAL OF EXPERIMENTAL PSYCHOLOGY-LEARNING MEMORY AND COGNITION	JEP:LMC	123
JOURNAL OF MEMORY AND LANGUAGE	J Mem Lang	150
JOURNAL OF PHONETICS	J Phon	47
JOURNAL OF PSYCHOLINGUISTIC RESEARCH	J Psycholing Res	28
JOURNAL OF SPEECH LANGUAGE AND HEARING RESEARCH	J SLHR	26
JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA	JASA	52
LANGUAGE AND SPEECH	Lang & Speech	28
LANGUAGE COGNITION AND NEUROSCIENCE	Lang Cogn Neuro	120
LANGUAGE LEARNING	Lang Learn	28
LINGUA	Lingua	28
MEMORY & COGNITION	M&C	43
NEUROIMAGE	Neuroimage	27
NEUROPSYCHOLOGIA	Neuropsychologia	37
PLOS ONE	Plos One	144
PSYCHOLOGICAL SCIENCE	Psych Sci	25
PSYCHONOMIC BULLETIN & REVIEW	PBR	48
QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY	QJEP	84
READING AND WRITING	Read Writ	27
SCIENTIFIC REPORTS	Sci Rep	32

labeled with the level that is being compared to the grand mean. A plot of the modeled data transformed into proportions can be found in Fig. 2; in this plot, the grey horizontal line reflects the model intercept (grand mean), and each point reflects the estimate for a particular journal.

Four journals are reliably better than average: these are *Bilingualism, Language, and Cognition*, *Journal of Phonetics*, *Journal of Memory and Language*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Two journals are reliably worse: *Frontiers in Psychology* and *Frontiers in Human Neuroscience*.

The differences between journals suggest that there is a crucial role for journals—and the editors and reviewers that contribute to the review process—in how models are reported and whether in-depth contrast description is encouraged or overlooked. We would like to applaud the journals that are at the top of best-practices in this domain and the individuals that have helped make this happen.

Patterns by keyword are varied

In order to examine the role of topic-specific conventions, we next examined the keywords associated with contrast description. We used the spell-check procedure in MS Excel to correct any misspellings and to Americanize all words, and replaced all punctuation and spaces with '_' in order to collapse similar terms for analysis e.g. *eye-tracking* and *eye tracking*. After doing this, there were 6758 unique keywords associated with 2553 unique papers.

We filtered the full data set (including all journals) for the 29 keywords with at least 30 observations each. We selected 30 as our cutoff because we desired to use a model with a more complex random effect structure than the by-journal analysis; requiring a larger sample size per keyword helps avoid any convergence issues. These data were submitted to a generalized linear mixed model with a random intercept for the journal that the keyword appeared in and for the year of publication. In this analysis, the factor keywords was again sum-coded, with the median level 'language production' as the reference (the omitted level). This again allows us to test whether each keyword is associated with a reliably different outcome than the grand mean of all keywords. Results of this model can be found in Table 4, and plot of the modeled data transformed into proportions can be found in Fig. 3; again, the grey horizontal line reflects the model intercept (grand mean), and each point

Table 3
By-journal contrast description analysis. Model intercept reflects the average level of contrast description and each effect reflects whether a journal is reliably different from average. Reference (omitted) level is 'Cognitive Science'. Model formula is: glmer(ContrastsUse ~ Journal + (1|Year), family='binomial')

Fixed Effects				
	Est.	S.E.	z val.	p
(Intercept)	−0.71	0.10	−6.79	0.00
B:L&C	1.05	0.31	3.40	0.00
JEP:LMC	0.67	0.19	3.56	0.00
Cog Psychol	0.65	0.44	1.47	0.14
J Phon	0.65	0.29	2.21	0.03
J Mem Lang	0.60	0.17	3.42	0.00
J Child Lang	0.42	0.38	1.12	0.26
Lang Learn	0.47	0.37	1.24	0.21
Lingua	0.51	0.38	1.36	0.17
AP&P	0.31	0.34	0.90	0.37
JEP:HPP	0.22	0.30	0.74	0.46
Lang & Speech	0.22	0.38	0.58	0.56
J SLHR	0.12	0.40	0.31	0.76
Lang Cogn Neuro	0.19	0.19	0.95	0.34
Sci Rep	−0.03	0.37	−0.10	0.92
Cognit	0.12	0.21	0.56	0.58
Neuroimage	0.10	0.39	0.26	0.79
JASA	0.11	0.29	0.37	0.71
QJEP	−0.01	0.23	−0.05	0.96
J Exp Child Psychol	−0.15	0.43	−0.34	0.73
Read Writ	−0.08	0.40	−0.20	0.84
Psych Sci	−0.06	0.42	−0.13	0.90
Plos One	−0.11	0.19	−0.58	0.56
PBR	−0.14	0.31	−0.46	0.65
JEP:G	−0.19	0.40	−0.47	0.64
Acta Psychol	−0.28	0.32	−0.88	0.38
J Psycholing Res	−0.42	0.41	−1.02	0.31
J Cog Psych	−0.45	0.44	−1.04	0.30
Appl Psycholinguist	−0.41	0.41	−1.00	0.32
Neuropsychologia	−0.52	0.38	−1.38	0.17
M&C	−0.52	0.36	−1.46	0.14
Front Psychol	−0.63	0.19	−3.33	0.00
Brain Lang	−0.84	0.54	−1.56	0.12
Front Hum Neurosci	−1.52	0.60	−2.55	0.01
Random Effects				
Group	Parameter		Std. Dev.	
Year	(Intercept)		0.23	
Grouping Variables				
Group	# groups		ICC	
Year	10		0.02	

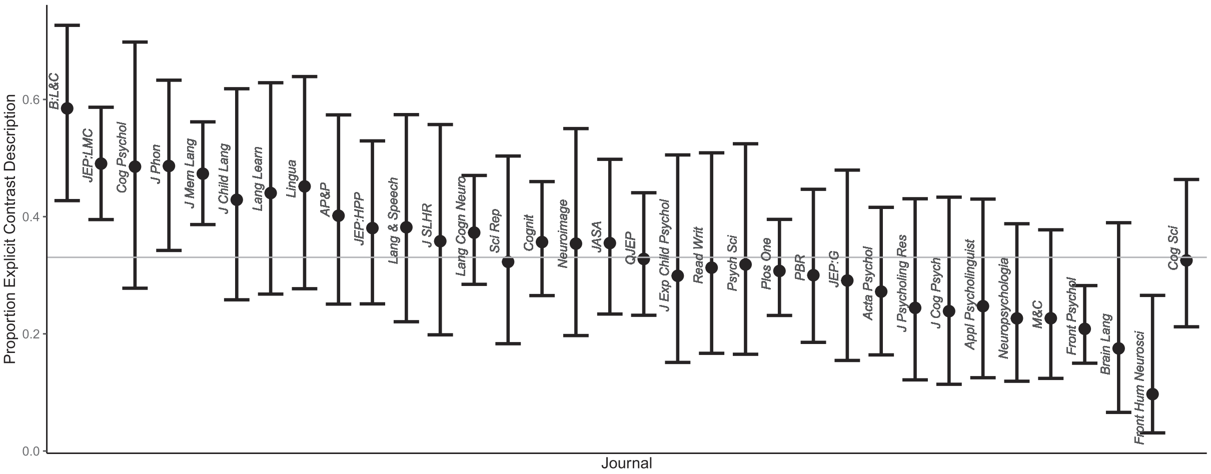

Fig. 2. Results from by-journal analysis, back-transformed into proportions.

Table 4
By-keyword contrast description analysis. Model intercept reflects the average level of contrast description and each effect reflects whether a keyword is reliably different from average. Reference (omitted) level is ‘language production’. Model formula is: glmer(ContrastsUse ∼ Keyword + (1 |Journal) + (1|Year), family=‘binomial’).

Fixed Effects

	Est.	S.E.	z val.	p
(Intercept)	−0.44	0.15	−2.86	0.00
structural_priming	1.04	0.36	2.89	0.00
individual_differences	0.77	0.32	2.41	0.02
eye_tracking	0.55	0.23	2.41	0.02
morphology	0.63	0.36	1.74	0.08
speech_perception	0.34	0.29	1.16	0.25
spoken_word_recognition	0.26	0.32	0.82	0.41
reading	0.29	0.20	1.51	0.13
sentence_processing	0.42	0.26	1.61	0.11
priming	0.25	0.34	0.72	0.47
working_memory	0.49	0.30	1.65	0.10
eye_movements	0.04	0.20	0.19	0.85
prediction	0.13	0.37	0.34	0.74
attention	−0.08	0.34	−0.23	0.82
bilingualism	−0.01	0.26	−0.03	0.97
syntax	0.12	0.34	0.34	0.73
memory	0.04	0.33	0.12	0.90
language_comprehension	−0.12	0.31	−0.39	0.70
speech_production	−0.23	0.37	−0.63	0.53
psycholinguistics	−0.36	0.39	−0.90	0.37
language_acquisition	−0.17	0.38	−0.44	0.66
lexical_access	−0.21	0.36	−0.59	0.55
emotion	−0.36	0.38	−0.95	0.34
word_recognition	−0.45	0.35	−1.27	0.20
masked_priming	−0.52	0.36	−1.45	0.15
prosody	−0.43	0.32	−1.33	0.18
language	−0.89	0.43	−2.07	0.04
visual_word_recognition	−0.87	0.38	−2.29	0.02
lexical_decision	−0.92	0.41	−2.21	0.03

Random Effects

Group	Parameter	Std. Dev.
Journal	(Intercept)	0.83
Year	(Intercept)	0.28

Grouping Variables

Group	# groups	ICC
Journal	161	0.17
Year	10	0.02

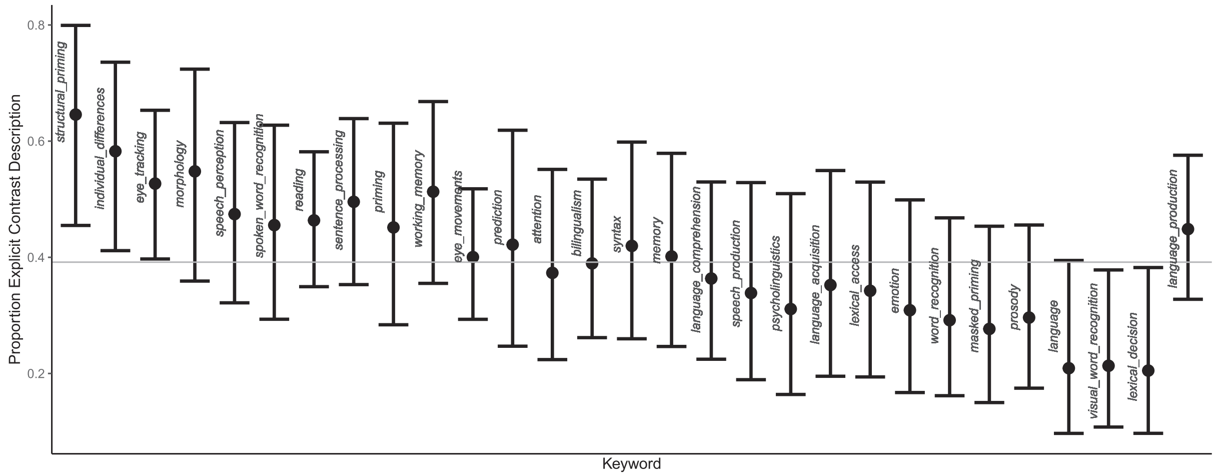

Fig. 3. Results from by-keyword analysis, back-transformed into proportions. Note that visual comparison does not necessarily reflect significance level because of the model random effects.

Table 5
Rates per keyword of papers that do not transparently describe contrasts, and contain at least one analysis with a significant interaction and significant main effect.

Keyword	Proportion Problematic Cases
attention	0.279
bilingualism	0.375
emotion	0.368
eye_movements	0.311
eye_tracking	0.272
individual_differences	0.217
language	0.344
language_acquisition	0.412
language_comprehension	0.333
language_production	0.357
lexical_access	0.351
lexical_decision	0.371
masked_priming	0.447
memory	0.429
morphology	0.343
prediction	0.250
priming	0.256
prosody	0.431
psycholinguistics	0.419
reading	0.281
sentence_processing	0.242
speech_perception	0.255
speech_production	0.286
spoken_word_recognition	0.341
structural_priming	0.158
syntax	0.366
visual_word_recognition	0.357
word_recognition	0.462
working_memory	0.269

reflects the estimate for a particular keyword.

Examining differences by keyword, while controlling for differences by journal and year, reveals a few differences by journal topic. Three keywords, *structural priming*, *eye tracking* and *individual differences* are better than the average, and three, *language*, *visual word recognition* and *lexical decision* are reliably worse. This suggests that idiosyncratic differences between fields matter: a few discrete areas have conventionalized on reporting contrasts, but most have not. While there are some bright spots, the implication is that any distortion of results due to a misunderstanding of contrast coding is spread across much of the field. In other words: a lack of understanding in contrast coding is likely a problem across most of psycholinguistics.

How much does it matter? Potentially, a lot

The analyses in this paper so far have focused on whether contrasts were sufficiently described for later reproducibility. In a strict and technical sense, models with unknown contrast coding schemes are uninterpretable: contrasts outline the hypotheses being tested when modeling, and not knowing these hypotheses means that the model should not be interpreted. Setting this aside, one can think about the inferences that would be licensed under various coding schemes. Here, it becomes clear that the wrong conclusions can be drawn about data when the researcher believes they are using sum coding but are actually using treatment coding. In this case, simple effects would be mistakenly interpreted as main effects. As demonstrated in the case study at the beginning of the paper, simple and main effects show different patterns in the presence of a reliable interaction: this suggests that analyses with significant main effects and significant interactions are places where model misinterpretation is especially likely.

To examine the rate of this type of mistaken inference in the literature, we performed a finer-grained coding of the 605 papers from the 'keywords' analysis that did not describe their contrasts. We chose this subset because it was of a tractable size and allowed us to examine further differences across subfields. For each of these papers, the first author coded whether any analysis in the paper included an interaction term, whether the interaction was significant, and whether any main effects were also significant.[6]

Of these 605 papers, 503 reported at least one analysis with an interaction term, and of those, 400 included a significant interaction. Of these 400 papers, 364 also report significant main effects. Under the hypothesis that when contrasts are not adequately described, authors typically use dummy coding but interpret results as sum coding, these 364 papers are highly likely to include false significant effects (type I errors). In other words: three-fifths of all analyses where contrasts are not reported meet the preconditions for a misinterpretation problem. Applying these proportions to the literature as a whole suggests that about 40% of the papers in the recent psycholinguistic literature are likely to contain one or more type I errors about a main effect.

Rates of papers meeting the preconditions for misinterpretation also vary by keyword, and this scales with rates of explicit contrast description by topic. These are reported in Table 5. Only 16% of papers with the keyword 'structural priming' are flagged as possibly erroneous, compared to 46% of papers with the keyword 'word recognition'. This suggests that while the contrast reporting problem occurs across the

[6] Thanks to Lotte Meteyard for inspiring this analysis.

field as a whole, it may have larger impacts on some sub-fields than others.

Discussion

We have presented evidence that the field of psycholinguistics does not provide sufficient detail about contrast coding for replicability, nor strictly speaking, for interpretability. Close to two-thirds of the over 3000 papers in our sample, regardless of the journal they appeared in or the research topic they focused on, did not describe their contrast use adequately. As we demonstrated in the case study presented above, failing to explicitly describe contrasts means that simple effects and main effects can be confounded with each other – if not by the author, then by the reader. This means that in the *majority* of the psycholinguistic literature sampled here, there are doubts about whether the reported effects can be replicated.

In this paper, we focused on coding the literature for whether contrast description was present in order to examine the boundary conditions of the replication problem. For the majority of the literature investigated, we did not assess whether reported results would be interpreted differently under different coding schemes, as we determined this was too time-consuming for a large sample and we believed that establishing the boundary conditions for the problem was most important. To get a more precise view on the problem, we then aimed to identify cases in a sub-sample of the data that were particularly likely to be problematic: papers containing at least one analysis with a significant interaction and at least one significant main effect. The rate of these cases is quite high, representing about 40% of all papers and approaching half of the literature in some domains. This implies that Type I errors about main effects are likely to be extremely common in the recent psycholinguistic literature.

Note however, some caveats about the magnitude of the problem. First, the same inferences can sometimes be made regardless of contrast coding choice. As we highlighted in the final analysis, it is important to remember that when simple and main effects show identical results (e.g., for main effects with no interaction), then confounding the two does not lead to an incorrect inference. Models with only one predictor will also always afford the same conclusions for sum and treatment coding. Similarly, the highest-level interaction in a model is invariant to contrast choice: if this term is the one for which the key predictions are made, correct conclusions will be made regardless of contrast coding. Finally, for models in which likelihood ratio testing is used to determine significance, contrast coding also makes much less difference, especially if Type II tests are run (where for any term, all of the effects it participates in are removed from the model; note that these are less popular than Type III tests).[7]

As such, it is certainly likely that of the sample reported here, many of the papers which did not report contrast coding did correctly interpret their conclusions – but given the low base rate of contrast reporting and the frequent use of study designs containing interactions in psycholinguistic studies, it is also likely that many false conclusions have been made, published, and cited over the past decade because of a misunderstanding of statistics. This means that we have established that many purported effects are impossible to replicate due to poor reporting and misinterpretation of contrasts, and have provided strong evidence that there is a fundamental problem in reporting and interpreting a now-standard statistical tool.

In our analysis, we showed three positive trends. Over time, contrast use has been increasing. This suggests that a deeper understanding of mixed models is being attained in our field, and that more transparent conventions are being adopted about model reporting. Literature on mixed modeling written for psychologists, such as Barr et al. (2013), Brown (2021), Matuschek et al. (2017), Meteyard and Davies (2020), Schad, Vasishth, Hohenstein, and Kliegl (2020), is likely contributing towards this upwards trend. We hope that this paper serves as part of a further change towards clear reporting of data analysis choices. Similarly, there is a role for journal-specific and topic-specific practices in explicit contrast description. This suggests that the influence of journal-specific practices, journal editors, and journal reviewers in particular topics has promoted behavioral change in the field. This is important, especially at the individual level: the review process should correct oversights in manuscripts in order to have the most rigorous, scientifically valid literature we can have. The downside of this fact is that when a paper appears in print with incorrect or opaque methodology, the authors and the reviewers may have not had a full understanding of the methods used. We hope that the tutorial presented above makes clear why it is important to specify contrast coding choices precisely, and point readers towards the textbook written by Winter (2019) and to the tutorial written by Schad et al. (2020) for more information. The UCLA Institute for Digital Research and Education has also written a document on contrast schemes in the R programming language that is quite approachable (UCLA IDRE, 2011) /.

We end with some recommendations for best practice regarding contrast coding. First, authors should in general, be able to describe and justify all choices made in analyzing data. This requires understanding the modeling procedure being used, rather than simply adopting the procedure that one 'should' use; however, note that even ANOVA models are more complex than they might seem on the surface. This means that we, as a field, may need to place more value in providing statistical training to students, and in employing statistical consultants for researchers to rely on when in doubt. We also suggest that it is better to use a tool that is well understood than to default to a tool that is popular, and caution reviewers and editors not to unduly pressure researchers to use MEM instead of other suitable techniques.

Models should be reported in full, including all fixed and all random effects, where present, and the choices made in selecting random effect structures, where present, should be clearly described in text (see Meteyard & Davies, 2020, for a comprehensive and clear set of guidelines for reporting models).[8]

When using a regression model (including an MEM), authors should clearly describe the contrasts associated with any categorical predictors, even if using the default treatment coding scheme, by either naming the

[7] We thank Dale Barr for pointing out rather important fine print on this statement. For Type III tests, where only the relevant term is dropped and not all associated higher level interactions, the choice of contrast coding does make a difference. In other words, comparing the full model `y ~a * b * c` to `y ~a * c` as a test of `b` is invariant to contrast coding, but comparing `y ~a * b * c` to `y ~a * c + a:b + b:c` (where : denotes an interaction without accompanying lower-level terms) is sensitive to choice of contrast, because the contrast determines the meaning of the `b`-terms left in the model. These types of "paradoxes" are part of the reason why Type-III tests are viewed as problematic (Venables, 1998).

[8] We should note we disagree with Meteyard and Davies on a few points. Their cited forum communication from Douglas Bates (Bates, 2006) states an opinion that has evolved over time: he currently argues that R2-like measures are problematic for mixed models and should probably not be used (Bates, personal communication). Likewise, the interpretation of correlation parameters in mixed models is problematic because a large number of groups (e.g., subjects or items) are required, because correlation estimates require a large number of samples before they stabilize (Schönbrodt & Perugini, 2013) and because the relevant sample size for the random effects is the number of groups, not the number of observations within them; nonetheless, given that they are part of the model's output, it may still be advisable to report them, though not to interpret them. Finally, we believe the notion of "convergence" was not sufficiently handled because lme4 tends to also issue convergence warnings for singular models, even when those models have converged, since the gradient-based convergence test is not valid for singular models. Nonetheless, we agree that because singular models are indicative of overfitting and present other inferential difficulties, it is often prudent to avoid them.

coding scheme or specifying the contrast matrix (e.g. *Factor A (magenta, green) was treatment coded* or *the three levels of Factor B, coffee, tea, and cocoa, were coded with two contrasts: (.25,.25, -.5) and (.5, -.5.0)*). Authors should also paraphrase what comparisons the contrasts make for easy interpretation of results by novices (e.g., *The model intercept therefore reflects the reference level of factor A, magenta* or *The first contrast tests caffeinated versus non-caffienated beverages, and the second tests coffee versus tea*), as we have aimed to do throughout this paper. Providing these two pieces of information, in text or in the caption to a model table, safeguards against the issues presented in the metascientific study above. A convention of interpreting contrasts directly also makes clear how the careful setting of contrasts eliminates most need for post hoc testing; additional post hoc tests (e.g., via `emmeans` in R) could still be done if necessary. If so, these should be clearly documented in the text (e.g. *An additional set of pairwise comparisons was performed to directly compare tea versus cocoa using the R package* `emmeans`.).

Finally, open science practices such as code and data sharing currently act as a last safeguard, allowing a dedicated reader to answer the question themself: we believe the results presented here emphasize the importance of open materials and especially, open data. We believe that the data are what is truly most important: the code and the data are the actual research, and publications are only the advertisement for it.[9] As such, the research product itself (code and data) should be made freely available and openly examinable and the associated advertisement (publication) should commit to full disclosure and truth in advertising (e.g., the full and transparent reporting of model structure and modeling decisions). However, this is the case only with one final caveat: The code that is used to conduct the analysis is per definition completely unambiguous, *as long as full version information is provided* (Simonsohn, 2021). As such, we recommend that authors use an appropriate environment tracker (e.g. `renv`, `groundhog`, `packrat` in R) to track versions and use software features for full-version reporting (e.g. `sessionInfo()` in R).

Conclusion

In 2008, a new method was presented to the field of psycholinguistics in a sufficiently compelling way that it became effectively mandatory to use mixed models in papers. However, the current results show that this change in analysis strategies was made without a full understanding of its implications. This means that as a field, we need to learn our methods better, and we need to be more cautious about ensuring we use methods that we understand. This suggests the further importance of methods training for researchers, especially when new tools emerge in the field. It also suggests that the field should in some cases be less dogmatic about the use of certain tools: while we believe that the virtues of MEM make it a method worth learning and understanding in full, it is not the drop-in replacement for ANOVA that some believe it to be, and should be properly understood before it is used. We hope that this paper increases the field's understanding of MEM, and we hope that it serves as a cautionary tale for what can happen with future adoption of new methods.

CRediT authorship contribution statement

Laurel Brehm: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Project administration. **Phillip M. Alday:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Thanks to the attendees of AMLaP 2020 for feedback and useful discussion. Thanks to Birgit Knudsen for assistance with data coding, and thanks to Dennis Joosen, Carlijn van Herpt, Inge Pasman, and Esther de Kerf for assistance with database prep. Thanks to Antje Meyer for encouraging this project to happen and for providing manuscript feedback, and thanks to Scott Fraundorf for teaching Laurel and the rest of the 2010–2012 UIUC Mixed Models Reading Group about contrast coding. We also wish to thank Dale Barr and Lotte Meteyard for extensive feedback on previous versions of this manuscript, especially on how to make its intended message more apparent and broadly accessible.

References

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.

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.

Bates, D.M. (2006). *[R] lmer, p-values and all that*. Post on the R-help mailing list, May 19th, available at: https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html Last Retrieved 2021-12-12.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bolker. B. (2021).*GLMM FAQ*. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html Last Retrieved 2021-12-12.

Breen, M. (2018). Effects of metric hierarchy and rhyme predictability on word duration in The Cat and the Hat. *Cognition, 174*, 71–81.

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245920960351

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (third ed.). Thousand Oaks, CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

He, J., Meyer, A. S., & Brehm, L. (2021). Concurrent listening affects speech planning and fluency: The roles of representational similarity and capacity limitation. *Language, Cognition and Neuroscience. Advance online publication*, 1–23. https://doi.org/10.1080/23273798.2021.1925130

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kretzschmar and Alday (to appear). Principles of statistical analyses: Old and new tools. In Grimaldi, M., Y. Shtyrov, & E. Brattico, (Eds.), *Language Electrified. Techniques, Methods, Applications, and Future Perspectives in the Neurophysiological Investigation of Language*. Springer. https://doi.org/10.31234/osf.io/nyj3k.

Long, J. A. (2020). jtools: Analysis and Presentation of Social Scientific Data. *R Package Version, 2*(1). https://cran.r-project.org/package=jtools.

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.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Onkelinx, T., (2017). *Using a variable both as a fixed and random effect* https://www.muscardinus.be/2017/08/fixed-and-random/.

Nieuwenhuis, R., te Grotenhuis, H. F., & Pelzer, B. J. (2017). Weighted effect coding for observational data with wec. *The R Journal, 9*(1), 477–485.

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects Models in S and S-Plus*, 3–56.

Venables, W.N. (1998). Exegeses on Linear Models. S-PLUS User's Conference.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*, 104038.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize. *Journal of Research in Personality, 47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Simonsohn, U. (2021). *Groundhog: Addressing The Threat That R Poses To Reproducible Research*. https://datacolada.org/95 Last Retrieved 2021-12-12.

Sweeney, R. E., & Ulveling, E. F. (1972). A transformation for simplifying the interpretation of coefficients of binary variables in regression analysis. *The American Statistician, 26*(5), 30–32.

[9] Thanks to Dale Barr for this analogy.

UCLA IDRE (2011). *R Library contrast coding systems for categorical variables*. https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/. Last Retrieved 2021-12-12.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Zhu, H. (2021). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4/ https://CRAN.R-project.org/package=kableExtra.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.