Contrast coding choices in a decade of mixed models

Brehm & Alday, *JML* (2022)

Pratik Bhandari

Paper discussion (PLEAD @ vd-mit)

28.10.2022

Summary of the paper

- Different contrasts can lead to radically different inferences
- Insufficient reporting and/or misunderstanding of contrast coding scheme in psycholinguistics
  - Improving over the years; worse in some journals and sub-fields than others
- Challanges to reproducibility
- Well understood tool $>$ Popular tool

Historical context

- MEM replaced ANOVA after Baayen et al. (2008) and Jaeger (2008) in psycholinguistics
  - MEM is considered the default approach and ANOVA "archaic" (cf. Gelman, 2005)
- No explanation why they used treatment coding, which is the default in R
  - "*Here I have used treatment-coding, because it is the most common coding scheme in the regression literature.*" (Jaeger, 2008, p. 436)

Another most cited paper: Barr et al. (2013) on random effects structure (maximal model)

Novelties of MLM

- More tinkering with software required than the ANOVAs
- More choices to be made in every step, knowingly or unknowingly (more *forking paths*)
  - Fixed and random effect structures,
  - Contrast coding,
  - Model comparison and selection, etc.

Development

- Introductory to advanced textbooks and tutorials
  - Pinheiro & Bates, 2000; Gelman & Hill, 2006; Zuur et al., 2009; Winter, 2019; McElreath, 2020; Meteyard & Davies, 2020; Brown, 2021
- Much focus on 'how to use MLM' and justify model structure
- Only a few papers about contrast coding
  - Schad et al., 2020; Rabe et al., 2020

Case study

- Comparison of treatment (or dummy) coding and sum (or effect) coding
- Demonstrates that ignorance of contrast specification and misinterpretation can result in wrong inference

Highlight

"Contrast coding for one variable also changes the interpretation of other variables. 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."

Case study

Treatment coding

- Intercept is the reference level (0 of (0,1) coding)
- Comparisons are the differences between the reference and non-reference levels
- Zero is the factor level coded as reference in (0,1)

Sum coding

- Intercept is the grand mean, i.e., mean of all factor levels
- Comparisons are the differences between the grand mean and the non-reference level
- Zero is the average of two levels in (-1,1)

Case study

- A: *Utensils*

Treatment contrast:

```
       Spoon
Fork       0
Spoon      1
```

Sum contrast:

```
      [S.Fork]
Fork         1
Spoon       -1
```

- B: *Foods*

Treatment contrast:

```
       Soup
Salad     0
Soup      1
```

Sum contrast:

```
      [S.Salad]
Salad         1
Soup         -1
```

Case study

Model with treatment contrast

```
m1 <- lmer(RT ~ Utensils*Foods + (1|Participant) + (1|Item), data=ds)
```

```
                          Estimate Std. Error   t value
(Intercept)               4.809597  0.2051378  23.44569
UtensilsSpoon             5.061245  0.2009573  25.18568
FoodsSoup                 5.128478  0.2009573  25.52024
UtensilsSpoon:FoodsSoup -10.095654  0.2841965 -35.52350
```

All effects are interpreted at the reference level.

At the level of Fork for Utensils, and Salad for Foods.

- Main effects are, in fact, simple effects.

Case study

- `UtensilsSpoon` is Spoon minus Fork at Salad
- Diff. in eating time for spoon and fork while eating salad

Utensils	r
Fork	4.809597
Spoon	9.870843

- `FoodsSoup` is Soup minus Salad at Fork
- Diff. in eating time for soup and salad while eating with fork

```
# A tibble: 2 x 2
  Foods     r
  <fct> <dbl>
1 Salad  4.81
2 Soup   9.94
```

9.94-4.81 = 5.13

9.87-4.81 = 5.06

These numbers match with the beta estimates of `UtensilsSpoon` and `FoodsSoup` respectively.

Case study

Model with sum contrast

```
m2 <- lmer(RT ~ Utensils*Foods + (1|Participant) + (1|Item), data=ds2)
```

	Estimate	Std. Error	t value
(Intercept)	7.380545264	0.16412676	44.96856681
Utensils[S.Fork]	-0.006709237	0.07104912	-0.09443096
Foods[S.Salad]	-0.040325381	0.07104912	-0.56757043
Utensils[S.Fork]:Foods[S.Salad]	-2.523913419	0.07104912	-35.52349896

Here, the main effects are not significant (cf. *m1*).

Case study

The higher order interaction is generally invariant to the choice of contrast.

Significant `Utensils:Foods` – slower to eat salad with a spoon & soup with a fork

- *m1* (`UtensilsSpoon:FoodsSoup`): β=-10.1, t=-35.5
- *m2* (`Utensils[S.Fork]:Foods[S.Salad]`) : β=-2.5, t=-35.5

! About interactions

Doesn't an interaction represent the difference between differences, without informing which one of the two differences is significant or greater than the other, and without showing the direction of the differences?

Case study: Demonstrated problem

Misinterpretation of simple effects and main effects if contrast coding scheme is not known/specified.

(But see the cases of post-hoc tests, LRT, higher order interactions)

Metascientific study

How well the studies that cited the 2008 papers describe their contrast coding scheme?

Well enough for reconstruction / reproduction?

Change over time in the last decade

Differences across sub-domains and journals

Meta-analysis

- Only 1069 out of 3125 papers describe contrasts.
- 66% of the sampled papers do not clearly describe their contrast coding.
- Correct interpretation and inference can't be drawn.
- Unclear what hypotheses are tested
  - Especially in the presence of crossover interactions

Trend: Improvement over time

- Positive linear trend in contrast coding description

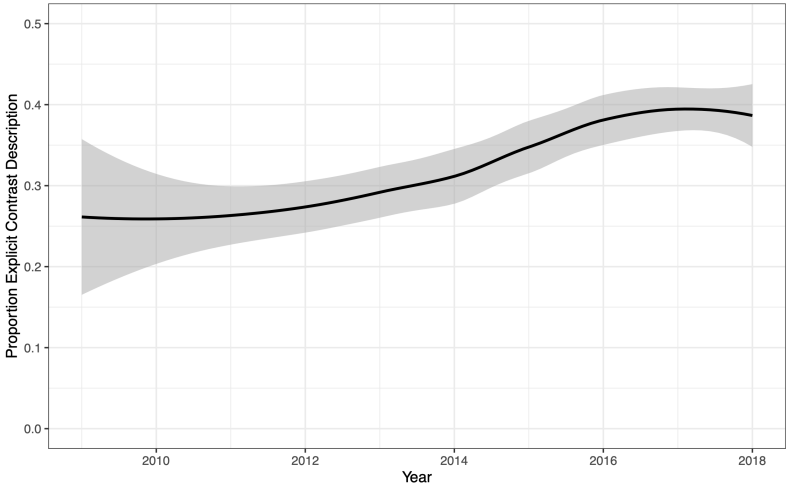

Figure 1: Proportion of explicit contrast use by year, with loess smooth.

Some journals are better than others

```
glmJournal <- glmer(ContrastsUse ~ J + (1|Year), family='binomial',
data=topP)
```

- *Sum coding*?
- *Simple coding* (stats.idre.ucla.edu; Vasishth et al., 2022)
  - Grand mean in the intercept
  - Median level: *Cognitive Science* as the reference

Some journals are better than others

- BL&C, JEP:LMC, J Phon, and JML are reliably above than average.
- Frontiers (Psychol and Hum Neurosci) is below average.

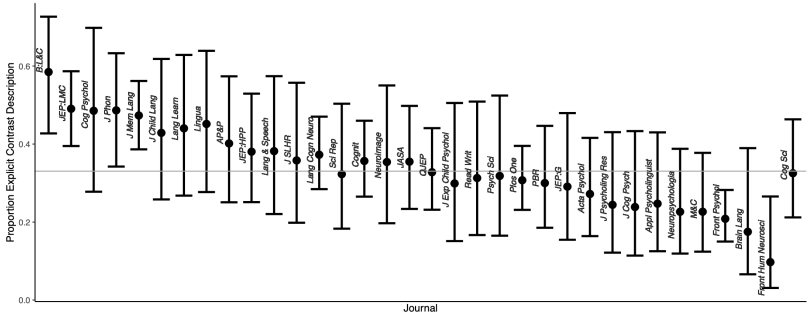

Figure 2: Proportion of explicit contrast use by journal.

Role of journals

Journals, editors, and reviewers in putting a check in model description

(also not to put an unnecessary pressure in using a less-understood tool like MLM)

! But...

What's the place of pre-prints and post-publication reviews in quality control?

Some sub-fields are better than others

6758 unique keywords from 2553 unique papers

```
glmKey <- glmer(ContrastsUse ~ K + (1 |Journal) + (1|Year), ...
```

- Contrast coding same as earlier
  - Grand mean in the intercept
  - Median level: *language production* as the reference

Some sub-fields are better than others

- *structural priming*, *ID*, and *eye-tracking* are reliably above than average.
- *visual word recognition* and *lexical decision* are below average.

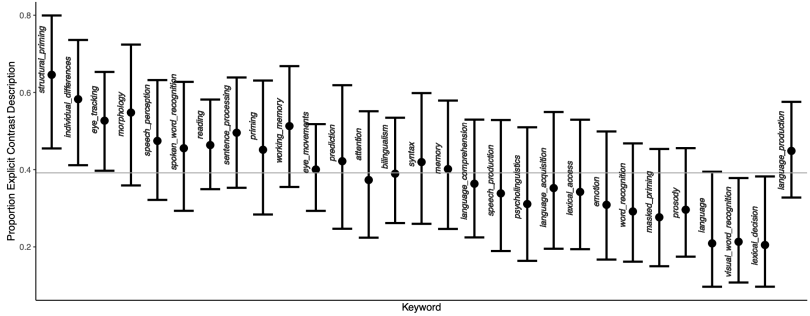

Figure 3: Proportion of explicit contrast use by keyword/sub-field

- *language comprehension* is worse than *language production* LOL

Rate of wrong contrast interpretation

No contrast description: 605 of 2553 papers

- Analysis with an interaction term: 503 of 605 paper
  - Significant interaction: 400 of 503 papers
    - Significant main effects: 364 of 400 paper

Rate of wrong contrast interpretation

Assumptions about 364 papers: - Used treatment coding - Interpreted as sum coding

364 of 605 papers (~60%): Type I errors

*Hinged on the assumption of treatment vs sum coding

Misinterpretation in sub-fields of psycholinguistics

$\lesssim$ 25% of the papers with these keywords are 'problematic':

- ID
- prediction
- sentence processing
- structural priming
- **NOT** eye-tracking

Misinterpretation in sub-fields of psycholinguistics

$\gtrsim$ 40% of the papers with these keywords are 'problematic':

- language acquisition
- masked priming
- memory
- prosody
- psycholinguistics
- word recognition
- **NOT** visual word recognition and lexical decision

Conclusion

- Confounding simple effects and main effects (cf. LRT, single predictor model, post-hoc tests, etc.)
- Type I error (on main effects) in almost half of the published sampled literature
- Contrast choice speaks to the hypothesis tested: impact on reproducibility
- Contrast specification: Some journals and some sub-fields are better than others
- More oversight from reviewers and journals necessary (?!)
- MLMs shouldn't be mindless replacements for ANOVAs (see Gelman, 2005)
- Gelman's wise words (?): Papers are ads for the research encased in data and code