Journal of Memory and Language 68 (2013) 255–278

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

Random effects structure for confirmatory hypothesis testing: Keep it maximal

Dale J. Barr [a,*], Roger Levy [b], Christoph Scheepers [a], Harry J. Tily [c]

[a] Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead St., Glasgow G12 8QB, United Kingdom
[b] Department of Linguistics, University of California at San Diego, La Jolla, CA 92093-0108, USA
[c] Department of Brain and Cognitive Sciences, Massachussetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 5 August 2011
revision received 30 October 2012
Available online 3 January 2013

Keywords:
Linear mixed-effects models
Generalization
Statistics
Monte Carlo simulation

ABSTRACT

Linear mixed-effects models (LMEMs) have become increasingly prominent in psycholinguistics and related areas. However, many researchers do not seem to appreciate how random effects structures affect the generalizability of an analysis. Here, we argue that researchers using LMEMs for confirmatory hypothesis testing should minimally adhere to the standards that have been in place for many decades. Through theoretical arguments and Monte Carlo simulation, we show that LMEMs generalize best when they include the maximal random effects structure *justified by the design*. The generalization performance of LMEMs including *data-driven* random effects structures strongly depends upon modeling criteria and sample size, yielding reasonable results on moderately-sized samples when conservative criteria are used, but with little or no power advantage over maximal models. Finally, random-intercepts-only LMEMs used on within-subjects and/or within-items data from populations where subjects and/or items vary in their sensitivity to experimental manipulations always generalize worse than separate F_1 and F_2 tests, and in many cases, even worse than F_1 alone. Maximal LMEMs should be the 'gold standard' for confirmatory hypothesis testing in psycholinguistics and beyond.

> *"I see no real alternative, in most confirmatory studies, to having a single main question—in which a question is specified by ALL of design, collection, monitoring, AND ANALYSIS."*
>
> Tukey (1980), "We Need Both Exploratory and Confirmatory" (p. 24, emphasis in original).

Introduction

The notion of *independent evidence* plays no less important a role in the assessment of scientific hypotheses than it does in everyday reasoning. Consider a pet-food manufacturer determining which of two new gourmet cat-food recipes to bring to market. The manufacturer has every interest in choosing the recipe that the average cat will eat the most of. Thus every day for a month (28 days) their expert, Dr. Nyan, feeds one recipe to a cat in the morning and the other recipe to a cat in the evening, counterbalancing which recipe is fed when and carefully measuring how much was eaten at each meal. At the end of the month Dr. Nyan calculates that recipes 1 and 2 were consumed to the tune of 92.9 ± 5.6 and 107.2 ± 6.1 (means ± *SD*s) grams per meal respectively. How confident can we be that recipe 2 is the better choice to bring to market? Without further information you might hazard the guess "somewhat confident", considering that one of the first statistical hypothesis tests typically taught, the unpaired *t*-test, gives $p = 0.09$ against the null hypothesis that choice of recipe does not matter. But now we tell you that only seven cats partici-

* Corresponding author. Fax: +44 (0)141 330 4606.
E-mail addresses: dale.barr@glasgow.ac.ukss (D.J. Barr), rlevy@ucsd.edu (R. Levy), christoph.scheepers@glasgow.ac.uk (C. Scheepers), hjt@mit.edu (H.J. Tily).

0749-596X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2012.11.001

pated in this test, one for each day of the week. How does this change your confidence in the superiority of recipe 2?

Let us first take a moment to consider precisely what it is about this new information that might drive us to change our analysis. The unpaired *t*-test is based on the assumption that all observations are *conditionally independent* of one another given the true underlying means of the two populations—here, the average amount a cat would consume of each recipe in a single meal. Since no two cats are likely to have identical dietary proclivities, multiple measurements from the same cat would violate this assumption. The correct characterization becomes that all observations are conditionally independent of one another given (a) the true palatability effect of recipe 1 versus recipe 2, together with (b) the dietary proclivities of each cat. This weaker conditional independence is a double-edged sword. On the one hand, it means that we have tested effectively fewer individuals than our 56 raw data points suggest, and this should weaken our confidence in generalizing the superiority of recipe 2 to the entire cat population. On the other hand, the fact that we have made multiple measurements for each cat holds out the prospect of factoring out each cat's idiosyncratic dietary proclivities as part of the analysis, and thereby improving the signal-to-noise ratio for inferences regarding each recipe's overall appeal. How we specify these idiosyncrasies can dramatically affect our conclusions. For example, we know that some cats have higher metabolisms and will tend to eat more at every meal than other cats. But we also know that each creature has its own palate, and even if the recipes were of similar overall quality, a given cat might happen to like one recipe more than the other. Indeed, accounting for idiosyncratic recipe preferences for each cat might lead to even weaker evidence for the superiority of recipe 2.

Situations such as these, where individual observations cluster together via association with a smaller set of entities, are ubiquitous in psycholinguistics and related fields—where the clusters are typically human participants and stimulus materials (i.e., items). Similar clustered-observation situations arise in other sciences, such as agriculture (plots in a field) and sociology (students in classrooms in schools in school-districts); hence accounting for the RANDOM EFFECTS of these entities has been an important part of the workhorse statistical analysis technique, the ANALYSIS OF VARIANCE, under the name MIXED-MODEL ANOVA, since the first half of the 20th century (Fisher, 1925; Scheffe, 1959). In experimental psychology, the prevailing standard for a long time has been to assume that individual participants may have idiosyncratic sensitivities to any experimental manipulation that may have an overall effect, so detecting a "fixed effect" of some manipulation must be done under the assumption of corresponding participant random effects for that manipulation as well. In our pet-food example, if there is a true effect of recipe—that is, if on average a new, previously unstudied cat will on average eat more of recipe 2 than of recipe 1—it should be detectable above and beyond the noise introduced by cat-specific recipe preferences, provided we have enough data. Technically speaking, the fixed effect is tested against an error term that captures the variability of the effect across individuals.

Standard practices for data-analysis in psycholinguistics and related areas fundamentally changed, however, after Clark (1973). In a nutshell, Clark (1973) argued that linguistic materials, just like experimental participants, have idiosyncrasies that need to be accounted for. Because in a typical psycholinguistic experiment, there are multiple observations for the same item (e.g., a given word or sentence), these idiosyncrasies break the conditional independence assumptions underlying mixed-model ANOVA, which treats experimental participant as the only random effect. Clark proposed the quasi-F (F') and min-F' statistics as approximations to an F-ratio whose distributional assumptions are satisfied even under what in contemporary parlance is called CROSSED random effects of participant and item (Baayen, Davidson, & Bates, 2008). Clark's paper helped drive the field toward a standard demanding evidence that experimental results generalized beyond the specific linguistic materials used—in other words, the so-called by-subjects F_1 mixed-model ANOVA was not enough. There was even a time where reporting of the min-F' statistic was made a standard for publication in the *Journal of Memory and Language*. However, acknowledging the widespread belief that min-F' is unduly conservative (see, e.g., Forster & Dickinson, 1976), significance of min-F' was never made a requirement for acceptance of a publication. Instead, the 'normal' convention continued to be that a result is considered likely to generalize if it passes $p < 0.05$ significance in both by-subjects (F_1) and by-items (F_2) ANOVAs. In the literature this criterion is called $F_1 \times F_2$ (e.g., Forster & Dickinson, 1976), which in this paper we use to denote the larger (less significant) of the two p values derived from F_1 and F_2 analyses.

Linear mixed-effects models (LMEMs)

Since Clark (1973), the biggest change in data analysis practices has been the introduction of methods for simultaneously modeling crossed participant and item effects in a single analysis, in what is variously called "hierarchical regression", "multi-level regression", or simply "mixed-effects models" (Baayen, 2008; Baayen et al., 2008; Gelman & Hill, 2007; Goldstein, 1995; Kliegl, 2007; Locker, Hoffman, & Bovaird, 2007; Pinheiro & Bates, 2000; Quené & van den Bergh, 2008; Snijders & Bosker, 1999b).[1] In this paper we refer to models of this class as *mixed-effects models*; when fixed effects, random effects, and trial-level noise contribute *linearly* to the dependent variable, and random effects and trial-level error are both normally distributed and independent for differing clusters or trials, it is a *linear mixed-effects model* (LMEM).

The ability of LMEMs to simultaneously handle crossed random effects, in addition to a number of other advantages (such as better handling of categorical data; see Dixon, 2008; Jaeger, 2008), has given them considerable momen-

[1] Despite the "mixed-effects models" nomenclature, traditional ANOVA approaches used in psycholinguistics have always used "mixed effects" in the sense of simultaneously estimating both fixed- and random-effects components of such a model. What is new about mixed effects models is their explicit estimation of the random-effects covariance matrix, which leads to considerably greater flexibility of application, including, as clearly indicated by the title of Baayen et al. (2008), the ability to handle the *crossing* of two or more types of random effects in a single analysis.