Statistics in Psycholinguistics: A critique of some current gold standards

R. Harald Baayen

Interfaculty Research Unit for Language and Speech, Max Planck Institute for Psycholinguistics, P.O.Box 310, 6500 AH, Nijmegen, The Netherlands.

Abstract

This paper presents a detailed critique of some current gold standards for the statistical analysis of experimental data in psycholinguistics. A series of examples illustrates (1) the disadvantages of reducing numerical variables to factors and the importance of including available covariates in the model, (2) the advantages of using multilevel models instead of the traditional by-subject and by-item procedures and the quasi-F test, and (3) the relevance of logistic models for binary data such as the error measure in decision tasks.

1 Introduction

The most commonly used statistical technique in psycholinguistics is analysis of variance. Generally, experimental research is planned in terms of factorial contrasts. Factorial designs are widely believed to be superior to multiple regression. Learning how to construct a data set with a factorial contrast while matching for a range of continuous predictors such as frequency of occurrence is regarded as an essential skill for experimental studies. As most psycholinguistic studies present a range of items to many different subjects, experimental data sets routinely undergo the averaging procedures of the by-subject and by-item analyses, applied indiscriminately not only to continuous variables such as response latencies, but also to dichotomous variables such as the accuracy measure. Many researchers seem to believe that the accepted

Email address: baayen@mpi.nl (R. Harald Baayen).

[1] This work was made possible by a PIONIER grant awarded by the Dutch National science foundation NWO. The author is indebted to Rob Schreuder, Mirjam Ernestus, and especially James Myers, for detailed comments and discussion.

statistical methods currently in use, and generally enforced by the journals, are the best that modern statistics has to offer.

The purpose of this study is to question the validity of this cluster of ideas and assumptions. It addresses two misconceptions in considerable detail: the overreliance on factorial designs when regression designs or analysis of covariance should be used (section 2), and the reliance on by-subject and by-item analyses of repeated measurement data where multilevel modeling provides a more insightful alternative (section 3). A third misconception is addressed in section 4, namely, that standard least squares analysis of variance would be appropriate for dichotomous response variables such as the accuracy measure.

2 The cost of dichotomization and factorization

Studies investigating language processing are faced with many numerical variables, both discrete (e.g., frequency of occurrence, word length in letters, sentence length in number of words, neighborhood size, age of acquisition) and continuous (e.g., word duration in ms, fundamental frequency in Hz). Nearly all studies addressing the potential role of such variables make use of factorial designs. In order to ascertain whether variable X codetermines processing independently of variables Y and Z, current practice is to carefully select words scoring either high or low on X while matching for Y and Z, pairwise, or in the mean for the high and low sets. Examples of studies using this methodology for word frequency, morphemic frequency, and syllable frequency effects in the mental lexicon, spanning some twenty years of reseach, are Taft (1979), Balota and Chumbley (1984), Levelt and Wheeldon (1994), Jescheniak and Levelt (1994), Sereno and Jongman (1995), Baayen, Dijkstra, and Schreuder (1997), Clahsen, Eisenbeiss, and Sonnenstuhl (1997), Hyönä and Pollatsek (1998), and Bertram, Schreuder, and Baayen (2000). It is widely believed that this is the most powerful means of ascertaining the independent effect of variables such as frequency of occurrence that are correlated with many other potentially relevant predictors.

Unfortunately, this belief is incorrect. Cohen (1983), in a paper entitled *The cost of dichotomization*, demonstrated that when a numerical predictor X is partitioned into a high versus a low group, for instance by creating a high and a low factorial contrast by splitting the data at the mean of X, this results in a substantial loss in power. Such dichotomization at the mean amounts to a degradation in the measurement of X. Precise numerical information on X is discarded in favor of a simple factorial contrast: 'high' versus 'low', or equivalently, 1 versus 0. Cohen showed that for bivariate normal distributions, dichotomization at the mean leads to a reduction in explained variance by a factor 0.637. More than one third of the variance that could have been explained using a regression analysis is left unexplained by the factorial analysis. This reduction in explained variance goes hand in hand with a reduction in power. Cohen shows that for a bivariate normal population with $\rho = 0.30$, a regression analysis using a sample of 80 cases has a probability of 0.78 of correctly rejecting the null hypothesis. Dichotomization reduces this probability to 0.57. A similar warning can be found in Harrell (2001), who states that

> Many researchers make the mistake of assuming that categorizing a continuous variable will result in less measurement error. This is a false assumption, for if a subject is placed in the wrong interval, this will be as much as a 100% error. Thus the magnitude of the error multiplied by the probability of an error is no better with categorization. (page 6)

To make this more concrete, consider a longitudinal language acquisition study in which the utterances of several children are recorded for 2 years at monthly intervals. Suppose that the number of passive forms in the recordings is the dependent variable of interest, and that the researcher is interested in the question whether passives are used more frequently as children become older. Dichotomazition at mean age would amount to replacing the actual ages of the children during the first year by the factor level "young" and their actual ages during the second year by the factor level "high". The counts for the 12 observations in each of the two levels of `Age` could then be compared using a t-test. The point made by Cohen (1983) is that a regression analysis in which number of passives is modeled directly as a function of `Age` is more powerful. In other words, when the values of a numerical predictor are available, it is disadvantageous to dichotomize such a predictor.

When designing experiments, a related question arises when it is known that the dependent variable Y is influenced by one or more correlated numerical predictors $X_1, X_2, \ldots$. Many studies opt for analysis of variance, and construct a factorial contrast in X_1 while matching for $X_2, X_3, \ldots$. The idea seems to be that by investigating an extreme contrast in X_1 while matching for the other variables, the possibility of detecting an effect of X_1 on Y is maximized. In what follows, we first consider the case in which there is one continuous predictor for Y, and then proceed to the case when there are two continuous predictors. I will use the term 'factorization' to denote the construction of a factorial contrast for extreme values of a predictor, to be distinguished from dichotomization, where a range of values is already available but split on mean or median and then assigned to 'high' and 'low' factor levels.

In the situation that there is a single relevant predictor X, building a factorial contrast for extreme values of X can be a useful strategy. An example is shown in the top left panel of Figure 1. The left panel shows an example of a dependent variable (Y) that is a linear function of X. We can construct a factorial contrast in X by assigning the first 5% of the ranked values of X

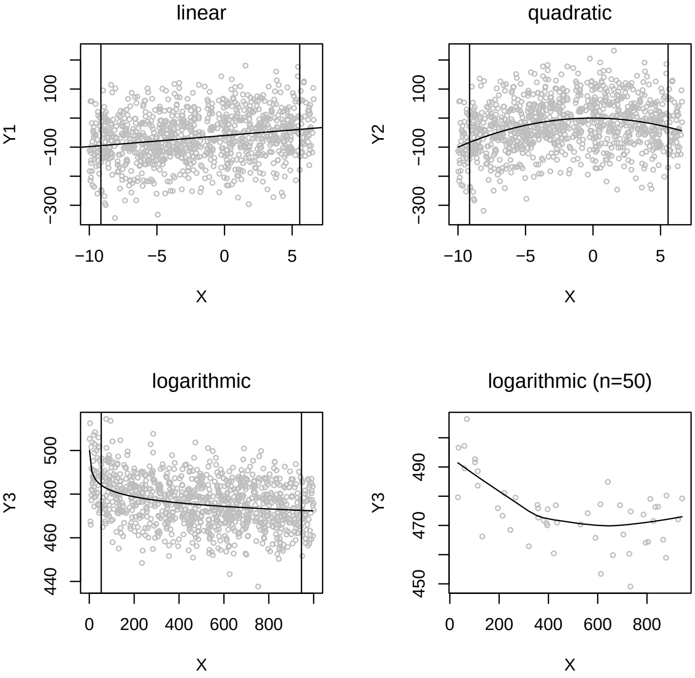

Figure 1. Contrasting the lowest 5% and the highest 5% of the values of X for a linear and a quadratic function of $X1 \in (-10, 6.6)$, and for a logarithmic function of $X2 \in (0, 1000)$.

to the low condition, and the last 5% of the ranked data points of X to the high condition. These two conditions are marked by the vertical lines in the upper left panel of Figure 1. The first row of Table 1 lists the proportion of simulation runs in which a t-test with 15 items from the low and 15 items from the high condition correctly suggests a significant effect ($p < 0.05$). It also lists the corresponding proportion for a regression analysis based on 30 randomly sampled items (see the appendix for further details). The factorial test clearly out-performs the regression by nearly a factor of two.

The high power of the factorization comes at a price, however. First of all, generalization is limited to the extreme ranges of X that were sampled. To see this, consider the upper right panel of Figure 1, which shows an example of a variable Y that is a quadratic function of X. The observations in the first and last 5% of the ranked data values of X are again assigned to the high and low conditions, which are marked by vertical lines, as before. The means for the low and high conditions are very similar to the means in the linear example in the upper left panel. It is important to note that in both cases, a factorial contrast will allow the researcher to assess whether there is an effect of X on Y, but not much else. No inference is possible about the remaining 90% of the data points. It is impossible to ascertain the nature of the relation of X and Y, whether it is linear or nonlinear. Even though a contrast between

Table 1
Proportion of 100 simulation runs in which a factorial t-test testing for differences in the mean of Y for fifteen values in each of the listed 5% ranges of X and a regression analysis based on a random selection of thirty values of X report a significant effect ($\alpha = 0.05$).

	5% ranges X		t-test	regression
	low	high		
linear function	1	20	0.49	0.22
quadratic function	1	20	0.41	0.40
logarithmic function	1	20	0.57	0.18
logarithmic function	2	19	0.26	0.18

the high and low groups may have been established, no prediction is possible for the values of X that fall outside the high and low conditions. The predicted mean values might fall in the interval bounded by the means of the high and low conditions, as in the upper left panel of Figure 1. However, the maximum value of Y might be reached for a non-extreme value of X, as illustrated for the quadratic function in the upper right panel.

A second disadvantage of factorization is that it need not be more powerful than regression. The second row of Table 1 illustrates this point for the quadratic example. For this example, the power of the (quadratic) regression analysis is very similar to that of the factorial analysis.

A third disadvantage of factorization is that the cutoff points for the low and high conditions are often arbitrary. Consider the lower left panel of Figure 1, which shows a logarithmic dependency on X. A factorial contrast based on the extreme deciles is again much more powerful than a regression analysis, as shown in the third row of Table 1, but in this case the high mean in the low condition is atypical compared to the neighboring data points. Researchers concerned about atypical values being observed for the extreme values of X might consider a factorial contrast on the basis of the second and nineteenth 5% ranges of X. For the logaritmic example of Figure 1, however, this more conservative procedure leads to a drastic reduction in power, as illustrated by the last line of Table 1.

The results for the regression analyses in Table 1 are based on the appropriate regression models (linear, quadratic, loglinear). The lower right panel of Figure 1 illustrates how non-parametric scatterplot smoothers such as the one developed by Cleveland (1979), see also Haerdle (1991), can bring nonlinearities to light and guide the formulation of the regression model, even for fairly small numbers of observations.

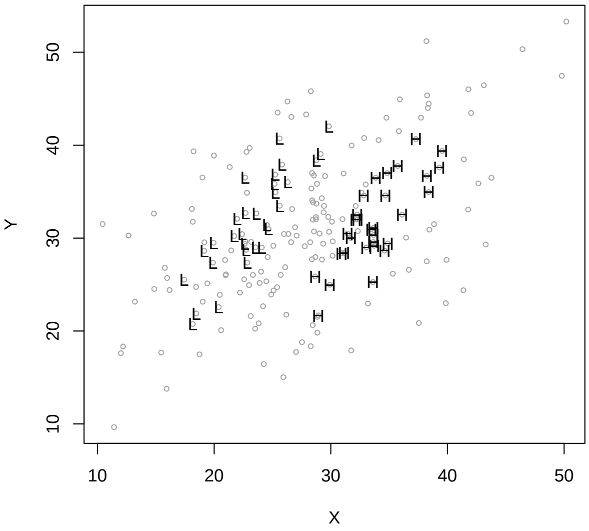

Figure 2. A simulated dataset with a factorial contrast in X matched in the mean for Y ($r(X, Y) = 0.50$).

In the light of these considerations, factorization is useful when obtaining data is costly and when documenting the existence of an effect is the sole purpose of the experiment, for instance, when the experiment is run to falsify some specific theory that crucially hinges on the presence or absence of an effect of a given predictor. It should be kept in mind, however, that regression may find non-linear relations in situations where factorial designs will never do so, for instance, when a factorial contrast is built around the extremes zero and π for the sine function on the interval $(0, \pi)$.

Thus far, we have considered the case in which there is one continuous predictor for the dependent variable. Next consider the case in which a dependent variable Z is a function of two continuous predictors, X and Y. In practise, the situation often arises that such predictors are correlated. For instance, word frequency, number of meanings, word length, mean bigram frequency, number of orthographic neighbors, and morphological family size are all correlated lexical variables (see, e.g., Baayen, Feldman, & Schreuder, 2003). Is it the case that factorization of X while matching for Y has the highest power for detecting an effect of X on Z? Many studies proceed along these lines. For instance, Taft (1977) and Bertram, Baayen, & Schreuder (1999), investigating frequency effects with visual lexical decision, factorially contrasted one frequency count (e.g., the frequency of the base word) while matching the other frequency count (e.g., the frequency of the complex word itself). Unfortunately, this procedure, instead of increasing power, may lead to a substantial decrease in power, as illustrated by the following simulation.

For each of 100 simulation runs, a data set was generated in which the dependent variable was a linear function of two correlated normally distributed predictors, X and Y, and an error term. From the 200 data points in a simulated data set, some 60 points were selected such that they were matched in the mean on Y. Half of these data points had a high value for X and half had a low value for X. Figure 2 shows an example of such a simulated data set. For each small band of Y values, the values of the high and low sets were at least 0.75 standard deviation of X apart. This reflects an often encountered situation in which, due to the correlation of X and Y, it is impossible to obtain sufficient items for factorial contrasts without overlapping intervals for the high and low conditions. Two sets of simulations were run, one set in which the beta coefficient of X was set to 3, and one in which this beta coefficient was set to zero. This made it possible to estimate power as well as type I error rate. Further details of this simulation are reported in the appendix.

Table 2
Power and type I error rate for a simulated data sets (100 runs), comparing a regression analysis based on a random sample of some 60 items, an analysis of variance of a factorial design with some 30 items in the high and low sets of X, and a regression analysis run on the data of the factorial design ($\alpha = 0.05$).

	power	type I error
regression	0.89	0.05
factorial regression	0.75	0.05
analysis of variance	0.40	0.01

For each simulated data set, three analyses were performed: an analysis of variance for the high versus low contrast of X, with in each set the same number of items (approximately 30, depending on the simulation run), a regression analysis based on the same items of the factorial design (labeled 'factorial regression' in Table 2), and a regression analysis based on a number of randomly selected data points equal to the number of items in the factorial analysis. Table 2 shows that the power of the regression analysis exceeds the power of the factorial analysis by a factor two. Inspection of the type I error rate shows, moreover, that in addition to a lack of power, the factorial design is too conservative as well — its Type I error rate equals 0.01 instead of 0.05 for $\alpha = 0.05$. Table 3 illustrates, furthermore, that the amount of variance explained by the factorial model is an order of magnitude smaller than that of the regression models. Note that the 'factorial regression', a 'post-hoc' regression run on the data points of the factorial design, has higher power and explanatory value than the factorial analysis.

Summing up, what this example illustrates is that it is counterproductive to try to achieve with a factorial analysis what should be done with regression. Instead of attempting to nullify the effect of covariates by means of matching

Table 3
Mean r^2 for the regression and factorial models for the simulated data sets of Table 2.

	present	absent
regression	0.60	0.46
factorial regression	0.49	0.34
analysis of variance	0.06	0.01

in the mean, the covariate should be brought straight into the model. This provides the regression model with the best means for assessing 'nuisance' variability, and for separating this variability from that due to the predictor variable of interest. The use of multiple regression, moreover, also allows the use of random samples, instead of the highly non-random samples of factorial studies. In addition, the practical problem of finding enough items under multiple matching constraints evaporates. Balota, Cortese, Sergent-Marshall, and Spieler (2003) discuss a number of additional reasons for not using factorial designs when regression is possible. Particularly noteworthy is their emphasis on the neglected role of explanation and predictive precision in psycholinguistics:

> . . . researchers should not be limited by the search for reliable effects and interactions, but also should attempt to determine how much variance a factor can account for. The primary driving force in the literature should no longer be if a variable has an impact on lexical processing, but also consider how much of an independent contribution the variable has on lexical processing. (page 9)

3 The cost of prior averaging

Inappropriate factorization and dichotomization is not the only practise in current psycholinguistics that involves the systematic loss of measurement information. The by-subject and by-item analyses that are currently the norm in psycholinguistic studies also bring along systematic data loss. It is widely believed that these averaging techniques are the best that current statistics has to offer. For instance, Raaijmakers, Schrijnemakers, & Gremmen (1999) reiterated the point made by Clark (1973), namely, that one common experimental design requires the quasi-F test. For another design, they recommend averaging by-subject. These recommendations build on what statistical theory had to offer in the 1940s and 1950s. However, these methods have various disadvantages that mathematical statisticians have addressed since then. The development of multilevel models, initiated in the 1970s (Lindley & Smith, 1972) has resulted in stable and well-studied algorithms that are now widely accepted in the statistical community (see, e.g., Bryk & Raudenbusch, 1992, Goldstein, 1995, Pinheiro & Bates, 2000, and Venables & Ripley, 2003). In what follows, I will introduce the main concepts of multilevel modeling, beginning with multilevel regression, followed by an example of a multilevel full factorial, and concluding with two more complicated designs, the Latin Square design discussed by Raaijmakers et al. (1999), and a design for which traditional analysis of variance would require a quasi-F test.

3.1 Multilevel regression

Consider a regression model in which the dependent variable, say RT, is a linear function of three predictor variables X, Y, and Z. Suppose that this model is tested for 20 items and that the experiment is run with 10 subjects. How might the results of this experiment be analysed?

One possibility is to calculate the mean RT for each item, averaging over the responses of the 10 subjects to that item. I will refer to this as the item regression. A second possibility is to run a regression analysis on the pooled data of all 10 subjects, without bringing the factor `Subject` explicitly in the model. I will refer to this as the 'simple regression' model. Simple regression should not be used, as observations from the same subject will in general not be independent. Consequently, the residual errors will not be independent but partially correlated, violating a basic assumption of regression and analysis of variance. A third alternative is discussed by Lorch and Myers (1990). They describe two equivalent models, of which the conceptually simpler one is known as random regression. In random regression, a separate regression model is fit to the data obtained for each individual subject. In the present example, we have 10 subjects, so 10 different regression models need to be fitted. Each regression model has four parameters: the intercept, and the coefficients for X, Y, and Z. In order to evaluate whether a predictor variable, say X, is significant, a t-test is performed on the 10 coefficients estimated for X.

Multilevel regression can be conceptualized as an extension of the random regression model. There are two important ways in which multilevel regression goes beyond random regression. The first is that only one model is fit to the data, instead of ten, in such a way that the fixed effects (the effects of X, Y, and Z) and the random effect (the `Subject` effect) are separated out on different levels. The fixed effects level of the model specifies how a unit change in one of the predictors affects the dependent variable when the other variables are held constant. The random effects level captures the variability associated with the subjects. Subjects generally differ with respect to their average response latencies. These differences are accounted for by means of a random variable with mean zero and unknown standard deviation. This unknown standard deviation is the (one and only) parameter in the multilevel model that accounts for the variability in the average response speed of the subjects. Given this parameter, estimates can be derived of the adjustments that have to be made to the intercept (specified in the fixed effect part of the model) such that the predictions for an individual subjects are as precise as possible. An important difference with a classical single-level regression model is that incorporation of subject as a factor in a classical (general linear) model requires a number of parameters equal to the number of subjects minus one, even though in theory a random effect is fully determined by its standard deviation. In multilevel regression, by contrast, only one parameter is required, the standard deviation of the subject random effect, as required. Thus, the multilevel model for n subjects and m items,

$$\begin{aligned} \mathrm{RT}_{ij} = & \overbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}}^{\text{fixed effects}} + \overbrace{b_j}^{\text{random effects}} + \epsilon_{ij} \\ & i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n, \\ & b_j \sim \mathrm{N}(0, \sigma_b^2),\ \epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2), \end{aligned} \tag{1}$$

has 6 parameters, the regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3$, and the standard deviations of the random effects, σ_b and σ_ϵ.

The second important difference between random regression and multilevel regression lies in the way in which the estimates of the coefficients are obtained. In random regression, the coefficients estimated for a given subject are exactly unbiased estimates of the true effects of the predictors when the model fits. However, the parameters derived from one dataset are not necessarily optimal for prediction to new datasets. The problem that arises for prediction is that the model will overfit the data. Typically, low predictions will be too low, and high predictions will be too high. In other words, the estimated parameters tend to shrink towards the mean in a new sample. This shrinkage is an adverse result of traditional modeling. The following simple simulation illustrates the problem. Consider an experiment with 10 subjects and 20 items, for which response latency RT is prediced from a single predictor X. Let

$$\begin{aligned} RT_{ij} = & \overbrace{400 + 5X_i}^{\text{fixed effects}} + \overbrace{b_j}^{\text{random effects}} + \epsilon_{ij} \\ & i = 1, 2, \ldots, 20;\ j = 1, 2, \ldots, 10, \\ & b_j \sim \mathrm{N}(0, \sigma_b^2),\ \ \epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2), \end{aligned} \tag{2}$$

where b_j (the subject random effect) and ϵ_{ij} (the residual error) are normally distributed random variables with zero mean and standard deviations $\sigma_b = 20$ and $\sigma_\epsilon = 50$ respectively. The left panel of Figure 3 plots the intercepts estimated for the different subjects in a random regression model, ordered from low to high. (A similar plot can be made for the estimated slopes.) The true

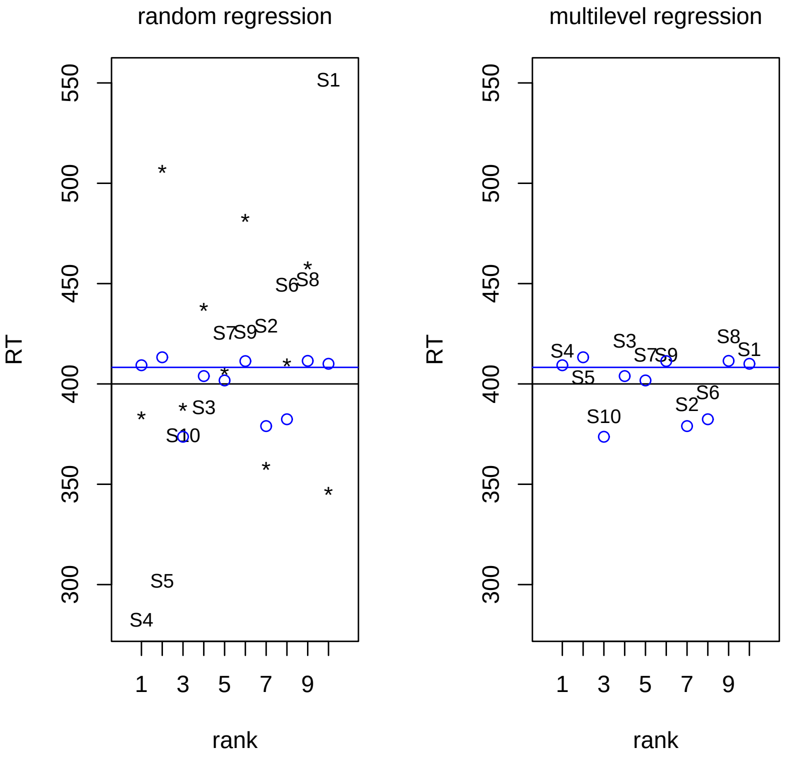

Figure 3. Ranked estimated random effects (labelled with subject numbers) and the true subject random effects (corresponding circles) in random regression (left panel) and in multilevel regression (right panel). In the multilevel regression, the estimates are shrunk in the direction of the mean. The horizontal lines denote the true mean (at 400) and its (slightly higher) sample estimate. The asterisks in the left panel represent estimated intercepts in a second experiment with the same subject random effect.

intercept β_0 is represented by a horizontal line ($RT = 400$). The circles represent the true subject effects, the b_j in (2). According to the random regression model, subjects S4 and S5 would have extremely low estimated intercepts, while subject S1 would have a very high estimated intercept, as shown by the labeled estimates. Although optimal in the least squares sense, these estimates are clearly way off, and in another experiment with the same subjects, the estimates will tend to regress towards the mean. This is illustrated by the asterisks, which represent a second experiment with the same subjects, and therefore with exactly the same random effect b_j but different residuals ϵ_{ij}. Note that the estimated intercepts for subjects S4 and S1 in this second experiment are closer to the mean, and that the estimate for S5 is again an outlier but this time in the opposite direction.

The right panel of Figure 3 graphs the estimated intercepts for the subjects in a multilevel regression model. The estimated intercepts are much closer

Table 4
Power and type I error for simulated RTs in multiple regression (100 simulation runs).

		Z present		
α	technique	X	Y	Z
0.05	multilevel regression	0.20	0.98	0.28
0.05	item regression	0.18	0.95	0.25
0.05	random regression	0.20	0.85	0.23
0.01	multilevel regression	0.10	0.83	0.12
0.01	item regression	0.06	0.75	0.08
0.01	random regression	0.04	0.68	0.09
		Z absent		
α	technique	X	Y	Z
0.05	multilevel regression	0.12	0.94	0.04
0.05	item regression	0.09	0.90	0.03
0.05	random regression	0.09	0.92	0.07
0.01	multilevel regression	0.03	0.87	0.01
0.01	item regression	0.00	0.70	0.01
0.01	random regression	0.03	0.69	0.02

to their true values than in the random regression model. What multilevel regression does, in other words, is to pre-shrink the estimates, bringing them closer to the true values and making more precise prediction possible.

Multilevel regression is not only preferable to random regression for its improved estimates of the random effects in the model, it is also somewhat more powerful, without giving rise to inflated Type I error rates. This is illustrated in Table 4. The upper half of this table lists the proportions of simulation runs (out of a total of 100 runs) in which the three predictors in a multiple regression model were correctly judged to be significant, for $\alpha = 0.05$ and for $\alpha = 0.01$. Details of this simulation can be found in the appendix. Note that the power of multilevel regression is at least as high as, and often higher, than the power of item regression and random regression. The lower half of Table 4 again reports the number of simulation runs in which a predictor was reported to be significant. This time, the underlying model had nonzero slopes for X and Y, but zero slope for Z, i.e., in this series of simulation runs, Z was not a predictor. As can be seen in the last column of Table 4, the proportion of runs in which Z is incorrectly judged to be significant is less than 0.05 when $\alpha = 0.05$ and 0.01 when $\alpha = 0.01$, as required. This simulation study illustrates the combination of slightly increased power and nominal Type I error rates that characterizes multilevel regression.

Thus far, we have only considered the simplest possible random effect structure in a multilevel model, namely, the random effect that accounts for differences among the subjects with respect to the intercept. However, subjects may also be differentially sensitive to predictor variables. Such interactions between subject and predictor variables can be modeled with great precision in multilevel modeling. The following example illustrates this for a real data set reported in Meeuwissen, Roelofs, and Levelt (in press). (I am indebted to Dr. Meeuwissen to making these data available to me). Naming latencies were obtained for digital clock times in Dutch, for a total of 20 subjects. The predictor of interest was the number of morphemes in the word to be named. This number of morphemes turned out to be a significant predictor in the expected direction: Naming latencies increased with the number of morphemes. Another experimental variable, however, turned out to be significant as well, the number of trials to which a subject had already responded in the experiment. As the number of trials increased, i.e., as subjects proceeded through the experiment, their response latencies decreased significantly.

At this point, the reader might wonder why one would want to include number of trials as a covariate. If the experiment was properly counterbalanced, shouldn't the effect of number of trials have been averaged out? The answer to this question is that counterbalancing guarantees that, in this example, the effect of number of morphemes is not confounded with the order in which the items appear in the experimental lists. Counterbalancing neutralizes bad side effects, but it does not account for the variance due to effects of habituation and effects of fatigue that might be present in the experiment. Bringing number of trials explicitly into the model formulation has three advantages. First, it enhances prediction accuracy. Second, since a greater proportion of the variance is accounted for, the residual error is smaller. As the residual error codetermines the standard error of the estimated coefficients, explaining more variance by bringing number of trials into the model enhances the probability of detecting a significant effect for number of morphemes. Third, explicit modeling of number of trials allows the researcher better insight in task-related effects in the experiment.

In the digital clock times naming experiment, a complex set of interactions emerged involving subject and number of trials. First of all, not all subjects evidenced the facilitatory main effect of number of trials. For some subjects, there was no observable effect of trial. In addition, there were two subjects for whom number of trials was in fact inhibitory. Figure 4 is a trellis graph visualizing this variability. Trellis plots are displays which contain one or more panels, arranged in a grid-like structure (a trellis), developed for data visualization by Cleveland (see, e.g., Becker, Cleveland, Shyu, & Kaluzny, 1995, and

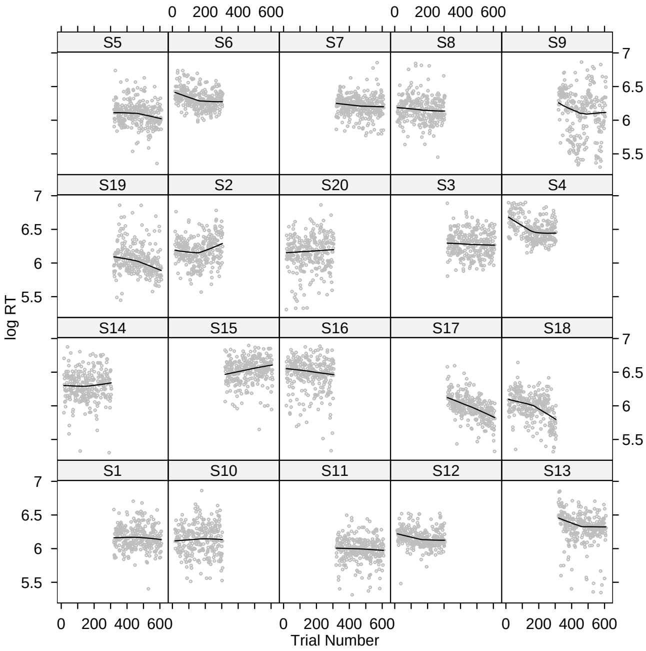

Figure 4. Trellis graph for digital clock naming times as a function of number of trials.

Becker, Cleveland, & Shyu, 1996). Trellis graphs often allow the researcher more insight into the structure of the data than formal statistical tests of some limited null hypothesis. Here, we use a trellis graph to obtain a visual summary of the relation between RT and trial number for the different subjects in the experiment. In this trellis graph, each panel represents a subject. The points in a given panel represent the items, placed in the plane spanned by trial number on the horizontal axis and by log RT on the vertical axis. The solid lines in the panels are loess non-parametric regression lines (Cleveland, 1979; see also Venables & Ripley, 1994, chapter 10). The digital clock times experiment was run jointly with another experiment, which explains why subjects are exposed to only early or only late trials. Note that number of trials tends to be roughly linear (further research might explore regression splines to account for potential nonlinearities, see, for instance, Harrell, 2001) and is negatively correlated with RT for a majority of subjects. This is in line with the significant main effect of number of trials in the multilevel model. However, there are subjects for which number of trials is not predictive (e.g., subject S10). For subject S15, number of trials is positively correlated with RT.

The model for this data set (with X denoting trial number and Y denoting number of morphemes, and with i ranging over items, j over subjects, and t over trials) is

$$\begin{aligned}\log \mathrm{RT}_{ij} = {} & \overbrace{6.2365 - 0.0002X_{t[ij])} + 0.0162Y_i}^{\text{fixed effects}} \;+\; \overbrace{b_j + b_{t[ij],j} + \epsilon_{ij}}^{\text{random effects}}, \qquad (3)\\ & i = 1, 2, \ldots, n;\; j = 1, 2, \ldots, 20;\; t = 16, 17, \ldots, 609;\\ & b_j \sim \mathrm{N}(0, 0.1568^2),\; b_{t[ij],j} \sim \mathrm{N}(0, 0.0003^2),\; \epsilon_{ij} \sim \mathrm{N}(0, 0.2020^2),\end{aligned}$$

where $t[ij]$ denotes the trial number for subject j responding to item i — due to counterbalancing, the trial number $t[ij]$ is different for each combination of `Subject` and `Item`. Note that we have a model with two random effects involving subjects. For each subject, we have an adjustment to the intercept (b_j), as well as an adjustment to the beta coefficient of the number of trials ($b_{t[ij],j}$). (For subject S15, for instance, this adjustment is large and positive, and reverses a negative slope into a positive slope.) Since there are two adjustments for a given subject, these adjustments might be correlated. Unlike random regression (or standard simple regression), multilevel regression provides the tools for investigating whether a parameter for the correlation between the adjustments for the intercept and the adjustments for number of trials needs to be included in the model. For the digital clock times naming data, such an extra parameter turned out to be significant: The adjustments (technically, the Best Linear Unbiased Predictors or BLUPs) for the intercept and the adjustments for number of trials were significantly negatively correlated ($r = -0.473$), as shown in Figure 5. Each circle in this plot represents a subject. Subjects with a greater positive BLUP for the intercept are the slower subjects. These are the subjects for which the BLUP for number of trials tends to be negative, for them, number of trials is more facilitating than for the subjects with negative BLUPs for the intercept. This random effects structure tells us something about how subjects performed the experiment, with slower, perhaps more careful subjects, gradually optimizing their performance as the experiment proceeded.

Summing up, multilevel regression yields more precise estimates, it has enhanced power combined with nominal type I error rates, and it allows more fine-grained control of the random effects structure in the model than traditional regression techniques such as random regression or item regression.

3.2 Multilevel analysis of variance

Analysis of variance can be understood as a special case of multiple regression (see, e.g., Chatterjee, Hadi, & Price, 2000, chapter 5) when dummy coding is used to represent factor levels. As in multilevel regression, fixed effects and random effects are separated out on different levels. In what follows, an example of a factorial multilevel model with a fairly complex random effect structure is presented first. Next, a Latin square design is analyzed. This section concludes with a comparison of multilevel modeling with quasi-F ratios

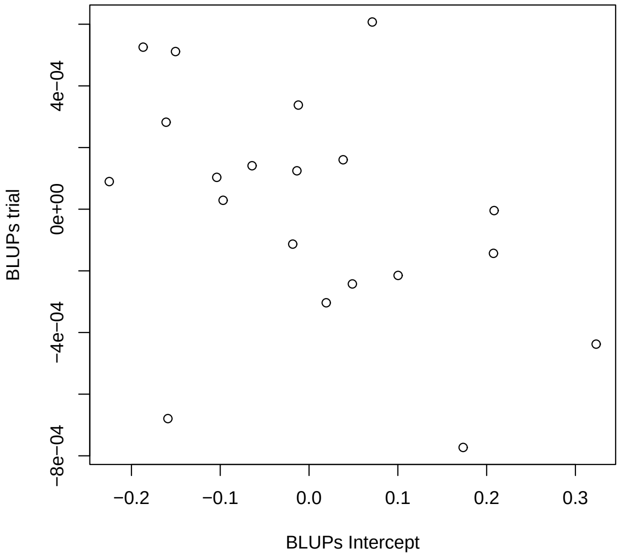

Figure 5. The negative correlation between the Best Linear Unbiased Predictors (BLUPs) for number of trials and the intercept in the digital clock times naming data.

and by-subject and by-item analyses.

3.2.1 An introductory example

In Taiwanese, tones that are normally realized with different contours of fundamental frequency (F0) may, in particular contexts, be realized with very similar F0 contours. Some theorists believe that this neutralization of the F0 contour is absolute. If so, neutralized tones should be indistinguishable to the hearer. However, it might also be the case that this neutralization is incomplete. If so, the F0 contours of neutralized tones would still be acoustically distinct. Myers and Tsay (2002) measured the F0 of three words (items) produced in the absence or presence of a listener by 17 speakers (subjects) at three points in the word (beginning, center, end) for two tones (yin, yang). If neutralization is absolute, these two tones should have indistinguishable fundamental frequencies. (I am indebted to Professor Myers for making these data available to me.)

There are three fixed effects in this design: `Tone` (with levels yin and yang), `Point` (with levels early, center, and late), and `Listener` (present versus absent). In addition, there are two sources of random variation, `Subject` and `Item`. The traditional procedure in psycholinguistics is to run two separate analyses of variance, one on means obtained by averaging over items, and one

Table 5
The 95% confidence intervals for the parameters of a multilevel model fit to the data on Taiwanese tone. sd() denotes standard error.

	lower	estimate	upper
Intercept	4.748	4.809	4.871
Tone (contrast yang: yin)	0.003	0.016	0.029
Point (contrast beginning: end)	-0.052	-0.040	-0.029
Point (contrast beginning: center)	-0.048	-0.037	-0.025
sd(Intercept)	0.081	0.120	0.177
sd(Tone)	0.009	0.015	0.026
sd(Listener)	0.020	0.030	0.044
sd(Point)	0.006	0.011	0.019
sd(Item)	0.045	0.058	0.075
sd(Residual Error)	0.042	0.044	0.047

on means obtained by averaging over subjects. These separate analyses are run because in classical analysis of variance this design does not allow the calculation of a unique F value for testing the effect of `Tone`. `Tone` might be tested against the interaction of `Tone` by `Subject`, or against the interaction of `Tone` by `Item`.

A multilevel analysis of variance obviates the need to run separate by-subject and by-item analyses. As in the previous example of multilevel regression, we separate the fixed effects from the random effects, with `Subject` as the main grouping factor for the random effects. By nesting the random effect of `Item` under `Subject`, we can account for `Item` effects as well as for possible interactions of `Subject` by `Item`. Table 5 lists the parameters of the multilevel model fit to these data, together with their 95% confidence intervals. `Tone` and `Point` emerged as significant main effects. The main effect of `Tone` suggests that tone neutralization might be incomplete.

It turned out that subjects were differentially sensitive not only to the items, but also to `Tone`, `Listener`, and `Point`. This was modeled by including additional random effects for these three fixed effects, all nested under subject, and all independent of each other and of the item random effect. Table 5 also lists the standard deviations of these random effects along with their 95% confidence intervals.

Figure 6 is a trellis graph plotting predicted versus observed log F0 for each subject in the experiment. There is one subject, S3, for which observed and expected F0 are uncorrelated, as shown by the zero slope of the nonparametric

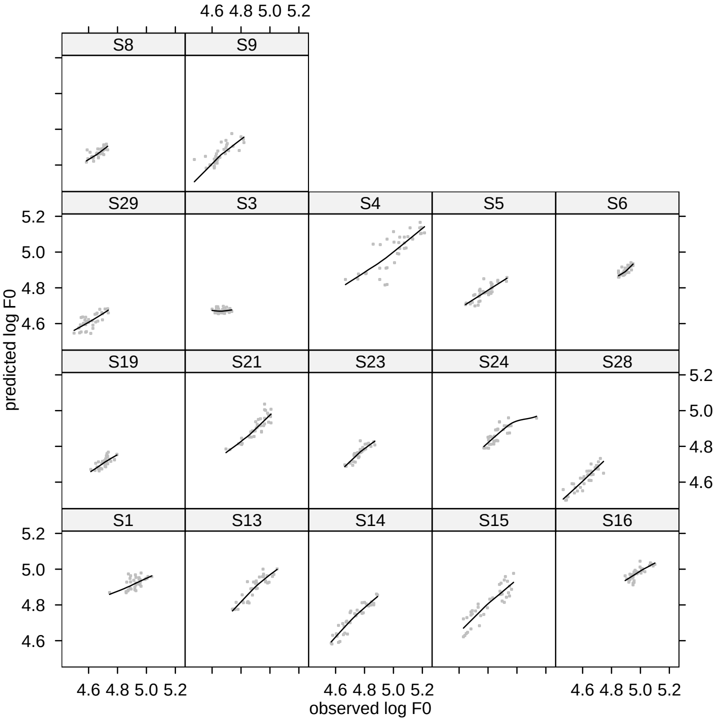

Figure 6. Observed and expected F0 for the data on Taiwanese tone neutralization.

regression line. For all other subjects, the data points cluster tightly around the nonparametric regression lines. Although there is some variation in the slopes, the predictions of the model are adequate, even when there is little variation in F0, as for subjects S6 and S8. The visual impression of a tight fit is supported by a high r^2 equal to 0.918. When the individual data points are predicted from a traditional by-subject analysis of variance model model, r^2 reduces to 0.763, indicating a loss of 15.5% of explained variance. The by-subject and the by-item analyses also underestimate the significance of the effect of `Tone`. The p-value for `Tone` is 0.0232 according to F1, and 0.0210 according to F2, while the multilevel model reports a p-value of 0.0105.

What is striking about the multilevel model is its parsimony in terms of number of parameters. The multilevel model has 10 parameters in all (listed in Table 5), whereas for instance the by-subject analysis requires 204 parameters (one for each data point after averaging over items) in order to test the effect of `Tone` against the interaction of `Tone` by `Subject`.

The next subsections discuss the possibilities of multilevel modeling for two simpler designs that received detailed attention in the study by Raaijmakers, Schrijnemakers, and Gremmen (1999).

3.2.2 A Latin Square design

Raaijmakers et al. (1999) discussed a simple Latin Square design for which they argued that the by-subject analysis would be appropriate. In their example, 12 items were presented at different stimulus onset asynchronies (SOA) to 12 subjects. The items were divided into three subsets, which were rotated across the three SOAs (short, medium, long). A given subject was exposed to exactly one presentation of each item, and each subset of four items was presented once in each of the SOA conditions.

The problem that this design poses is that a standard least squares decomposition does not allow the effect of SOA to be tested. The solution offered by Raaijmakers et al. is to average, for each subject, the four RTs in each of the three subsets. After this averaging process, the effect of SOA can be tested against the residual mean squares. For their data set, they report $F(2, 18) = 0.7781, p = 0.4741$.

Before addressing the question of what multilevel model might be fitted to this kind of data, it is useful to discuss the difference between crossing the random effect of `Item` with the random effect of `Subject`, and nesting `Item` under `Subject`. When considering items in relation to subjects, there are two strategies that can be followed. The first strategy is to assume that a given item will have exactly the same effect across all subjects. Since different subjects may have had different experience with different items, see, e.g., Gardner, Rothkopf, Lapan, and Lafferty (1987) and also Quené & van den Berg (2001), the assumption that a given item will have exactly the same effect across all subjects may be too strong. In order to allow differences between subjects with respect to a given item into the model, an interaction term for `Item` by `Subject` will therefore often be added to the model. This first strategy amounts to crossing items with subjects together with the interaction of these two random effects. Traditional analysis of variance with `Subject` and `Item` random effects proceeds from the assumption that `Subject` and `Item` are crossed.

The second strategy is not to commit oneself to the strong a-priori assumption that there should be a 'main effect' of `Item` across the subjects, but to proceed from the idea that the `Item` effect might be quite different for the individual subjects. This idea can be implemented by nesting `Item` under `Subject`. Nesting does not imply the necessary absence of a common effect of `Item` in the model. Such a common effect, if present, will be captured implicitly. Multilevel modeling presupposes this less restrictive assumption of items being nested under subjects.

When fitting a multilevel model to the abovementioned data from Raaijmakers et al. (1999), we begin with modeling `Subject` as the main grouping level of the random effects structure. Next, we include `Item` as a random effect nested under `Subject`, leading to the following model:

Table 6
Power and type 1 error for 100 simulation runs of the Latin Square design of Raaijmakers et al. (1999).

	Model with effect of `SOA`	
	Subject Analysis	Multilevel Analysis
alpha=05	0.42	0.50
alpha=01	0.25	0.29
	Model without effect of `SOA`	
	Subject Analysis	Multilevel Analysis
alpha=05	0.03	0.01
alpha=01	0.00	0.01

$$\begin{aligned}\mathrm{RT}_{i(j)k} = & \overbrace{\beta_0 + \mathrm{SOA}_k}^{\text{fixed effects}} \quad + \quad \overbrace{b_j + b_{i(j)}}^{\text{random effects}} \quad + \quad \epsilon_{i(j)k}, \\ & k = 1, 2, 3 \; i, j = 1, 2, \ldots, 12, \\ & b_j \sim \mathrm{N}(0, \sigma^2_{b_j}),\; b_i(j) \sim \mathrm{N}(0, \sigma^2_{b_i(j)}),\; \epsilon_{i(j)k} \sim \mathrm{N}(0, \sigma^2_\epsilon).\end{aligned} \tag{4}$$

The parameters of this model can be estimated, but this stretches the multilevel approach to its limits as for each data point two random effects have to be estimated. As a result, the confidence intervals for the standard deviations of the nested item random effect and the residual error are huge. When applied to the data from Raaijmakers et al. (1999), a p-value is obtained that is much more conservative: $F(2, 130) = 0.1057, p = 0.8998$.

With real experimental data, it is not known a-priori whether nesting or crossing is more appropriate. In the case of the present data, however, it is more likely that the simulation model that generated the data set crossed `Item` with `Subject`, as this is the default 'world view' underlying traditional analysis of variance. In order to accomodate `Item` as crossed with `Subject` in a multilevel model, we have to introduce it into the model as a fixed effect — multilevel models do not allow crossed random effects at the main grouping level. In what follows, I will first illustrate that this leads to the desired results by means of a simulation. The details of this simulation can be found in the appendix. Next, I will outline briefly why this is correct.

Table 6 lists power and Type 1 error rate for 100 simulation runs, comparing the by-subject analysis advocated by Raaijmakers et al. (1999) and a multilevel model with `Item` included as fixed effect crossed with `Subject`. Note that the power of the multilevel model is greater than that of the by-subject analysis, while at the same time its Type 1 error rate is in conformity with the nominal values.

Table 7
Power and Type 1 error for 100 simulation runs for the Latin Square design of Raaijmakers et al. (1999) combined with a longitudinal effect of fatigue.

	Model with effect of SOA	
	Subject Analysis	Multilevel Analysis
alpha=05	0.46	0.55
alpha=01	0.31	0.30
	Model without effect of SOA	
	Subject Analysis	Multilevel Analysis
alpha=05	0.43	0.04
alpha=01	0.23	0.00

Perhaps the greatest advantage of using a multilevel model is that longitudinal effects in the experiment, such as the effects of habituation and fatigue observed for the digital clock naming latencies in the data of Meeuwissen et al. (2003), can be brought into the model. Table 7 illustrates that the presence of such longitudinal effects can wreak havoc with the subject analysis. Table 7 reports power and Type 1 error rate when an effect of fatigue is built into the simulation. Even though there is some simple counterbalancing in the design (different subjects are exposed to different permutations of the subsets of items), the subject analysis has a fatally high Type 1 error rate. By contrast, a multilevel model including trial number as covariate combines similar power with an acceptable Type 1 error rate. Although with more extensive counterbalancing this adverse effect can be reduced for the subject analysis, it is only the multilevel analysis that can bring the effect of `Trial` directly into the model.

Having illustrated the advantages of multilevel modeling, we now return to the question why including `Item` as a fixed effect is appropriate when, as in this simulation, `Item` is truly crossed with `Subject` in the population. To answer this question, first note that modeling `Item` as fixed allows us to capture the variation due to the items, separating it from the residual error. This is important because in multilevel analysis of variance, as in regression, the standard errors of the coefficients are co-determined by the residual error. The smaller the residual error, the tighter the confidence interval of the coefficient, and the smaller the associated p-value will be.

Next, consider the definition of a random effect as a normally distributed random variable with zero mean and some unknown standard deviation. When we include `Item` as fixed in the multilevel model, when in fact it is random, three things happen. First, instead of having one parameter for the standard deviation of the item random effect, we have a number of parameters equal to

Table 8
Actual and estimated parameters for the multilevel model, estimates averaged over 100 simulation runs.

	simulation	estimate
standard deviation `Subject`	39	37.8
standard deviation `Item`	28	29.1
standard deviation of the error	20	20.1
intercept	534	534.1
contrast effect SOA short	5	5.2
contrast effect SOA medium	-4	-4.5
effect of fatigue	4	4.0

the number of items minus one. Second, instead of having estimates of the item adjustments (BLUPs) that are appropriately centered around zero, we have a series of item coefficients with, in general, a non-zero mean. The reason for this is the dummy coding of the factor `Item`. When contrast coding is used, for instance, the item coefficients represent contrasts between one specific 'pivotal' item that happens to be mapped onto the intercept (and hence has a zero coefficient) and each of the other items. The further the true adjustment of this pivotal item is from the mean of the item effect, i.e., the further away it is from zero, the larger the contrasts, and hence also the greater the absolute mean of these contrasts, will be. Third, if `Item` is a random effect, the fixed effect coefficients of `Item` will still be normally distributed with a standard deviation that will be an estimate of the true standard deviation of the `Item` random effect. This is illustrated in Table 8, which lists the estimated parameters and their true values in the simulation averaged over 100 simulation runs.

In other words, the only thing that is wrong with the multilevel model is that it cannot predict to new items while it should. It fails to do so only because the by-item adjustments are hard-wired into the model as fixed effects. We can adjust for this, fortunately, by centering the estimated item coefficients, not forgetting to including the zero coefficient for the pivotal item. This is illustrated for an arbitrary simulation run in Figure 7. The left panel plots the centered estimated effect of `Item` on the horizontal axis, and the true random effect of `Item` on the vertical axis. Note that there is a high correlation between the true item effects and their (centered) estimates. The right panel of Figure 7 is a quantile-quantile plot illustrating that the estimated coefficients of the `Item` effect are indeed normally distributed. This is supported by the high p-value of the Shapiro-Wilk test for normality ($p = 0.66$). Therefore, in order to obtain proper predictions for unseen, novel, items, we only need to adjust the intercept of our model, as the current intercept is specific to the pivotal item. We can do so by first adding, for each item, the value of its estimated coefficient to the intercept, followed by averaging. The resulting mean intercept is our best guess for the intercept for an item that was not in the experiment, and the centered estimated coefficients are now similar to the adjustments associated with a random effect in the multilevel framework.

In other words, even though we have included `Item` as a fixed effect in the model, we can still ascertain that it is in fact a random effect, and we can adjust the model so that we can generalize from our sample of items to the population of items. Thus, this example shows how one can go a route opposite to the one traditionally followed, by initially proceding from the assumption that the item effect is fixed instead of random, and subsequently relaxing that assumption upon inspection of the estimates obtained.

When the items in an experiment are not randomly selected from the population, for instance, because the researcher has screened the items or matched them carefully on a number of dimensions, the a-priori assumption that `Item` is a random effect may be unwarrented, especially in cases where a replication study would be hard set to obtain a second sample with a new, disjunct set of pairs. For the dangers inherent in non-random sampling of items, the reader is referred to Forster (2000). With respect to the present example, if `Item` were a fixed effect, this would have shown up as non-normality in the quantile-quantile plot of the item coefficients.

3.2.3 A comparison with the quasi-F test

Let's finally consider how multilevel modeling compares to classical analysis of variance with quasi-F ratios.

Suppose that total eye fixation durations are obtained for some region in 8 pairs of matched sentences that differ systematically with respect to some characteristic pertaining to linguistic complexity, henceforth `Treatment`. Let `Treatment` have two levels, simple versus complex, and assume that data are obtained for 6 subjects, with each subject reading all $2 * 8 = 16$ sentences. The question of interest is whether `Treatment` has an effect on total fixation durations. Table 9 list the outcome of a simulated experiment, and Table 10 lists the mean squares and the terms contributing to these mean squares in a standard analysis of variance decomposition (see, e.g., Clark, 1973, Raaijmakers et al., 1999, or Cobb, 1998, chapter 13). As there is no appropriate term (mean square) to test the effect of `Treatment` against — there is no term that differs from the term for `Treatment` in just one random effect — the textbook solution is to make use of a pseudo-F or quasi-F ratio F_C (Satterthwaite, 1946; Cochran, 1951) that isolates the `Treatment` effect by comparing sums of expected mean squares (EMS) that differ precisely with respect to the `Treatment` effect:

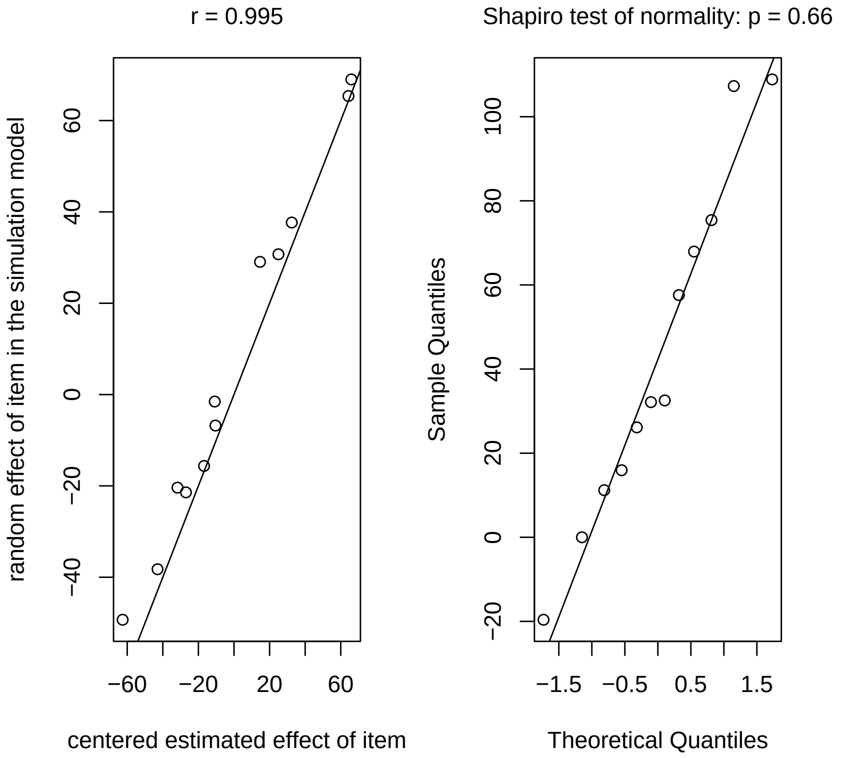

Figure 7. Centered estimated effect of `Item` (horizontal axis) and true random effect of `Item`. The right panel is a quantile-quantile plot of the centered estimated effect of `Item`.

$$
\begin{aligned}
F_C(\mathrm{df}_1, \mathrm{df}_2) &= \frac{EMS(T) + EMS(TSP)}{EMS(TS) + EMS(TP)} = \frac{(2E + 2TSP + TS + TP) + T}{(2E + 2TSP + TS + TP)} \\
\mathrm{df}_1 &= \frac{(TSP + T)^2}{TSP^2/\mathrm{df}(TSP) + (T^2/\mathrm{df}(T)} \\
\mathrm{df}_2 &= \frac{(TP + TS)^2}{TP^2/\mathrm{df}(TP) + TS^2/\mathrm{df}(TS)} \qquad (5)
\end{aligned}
$$

For this data set, Cochran's pseudo F ratio F_C equals 6.62, which, with 1.11 and 11.99 degrees of freedom, reaches significance ($p = 0.0221$). Clark (1973) and Raaijmakers, Schrijnemakers, & Gremmen (1999) argue that this would be the only appropriate statistical test.

However, tests involving pseudo F ratios are known to be very conservative, see, e.g., Forster and Dickinson (1976) and Wickens and Keppel (1983). The current gold standard in psycholinguistics is to carry out separate tests by subject and by item, as this procedure is somewhat less conservative than the test using the pseudo-F ratio. For the present data set, the combined F1 and F2 test also suggest that the effect of treatment is significant: $t1(5) = 3.9, p = 0.0115$ for the `Subject` means, $t2(7) = 3.3, p = 0.0137$ for the `Pair` means.

Table 9
A data set with `Treatment` as fixed effect and `Pair` and `Subject` as random effects.

	Treatment	Subj1	Subj2	Subj3	Subj4	Subj5	Subj6
Pair 1	simple	174.2	172.3	172.9	207.3	235.7	170.4
Pair 2	simple	207.2	184.7	206.2	181.9	185.9	227.5
Pair 3	simple	196.5	172.9	198.4	167.5	200.5	210.1
Pair 4	simple	205.4	222.7	205.6	207.5	227.8	246.4
Pair 5	simple	219.1	197.9	234.0	257.9	224.8	245.9
Pair 6	simple	249.6	240.8	218.8	279.9	210.3	231.6
Pair 7	simple	211.6	198.7	175.0	232.7	201.7	208.8
Pair 8	simple	226.5	208.5	197.6	212.1	231.3	234.4
Pair 1	complex	132.8	172.0	218.3	178.6	170.9	136.9
Pair 2	complex	193.0	146.3	173.4	166.6	171.9	220.9
Pair 3	complex	206.7	160.5	200.2	195.6	195.1	163.0
Pair 4	complex	150.5	200.5	195.2	182.5	183.0	201.7
Pair 5	complex	239.5	176.8	198.9	209.3	237.9	197.8
Pair 6	complex	192.7	166.3	221.5	212.3	205.2	212.8
Pair 7	complex	207.9	246.1	200.4	201.6	203.2	177.9
Pair 8	complex	250.9	182.1	217.2	229.6	201.9	225.8

Table 10
Decomposition for the data of Table 9 with `Treatment` as fixed effect and `Subject` and sentence `Pair` as random effects.

		Df	Sum Sq	Mean Sq	Decomposition
T:	Treatment	1	6755.0	6755.0	E+TSP+TS+TP+T
S:	Subject	5	3197.7	639.5	E+TSP+ST+SP+S
P:	Pair	7	21812.4	3116.1	E+TSP+TP+SP+P
TS:	Treatment:Subject	5	2226.3	445.3	E+TSP+TS
TP:	Treatment:Pair	7	4423.7	632.0	E+TSP+TP
SP:	Subject:Pair	35	19733.9	563.8	E+TSP+SP
TSP:	Treatment:Subject:Pair	35	13165.1	376.1	E+TSP
E:	Residuals	0	0.0		

Table 11
Least squares decomposition for a multilevel model for the data of Table 9.

	Df	Sum Sq	Mean Sq	F value	P-value
Error stratum: `Subject`					
Residuals	5	3197.7	639.5		
Error stratum: `Pair` nested under `Subject`					
`Pair`	7	21812.4	3116.1	5.5266	0.0002
Residuals	35	19733.9	563.8		
Error stratum: `Pair`					
`Treatment`	1	6755.0	6755.0	17.5552	0.0001
`Treatment:Pair`	7	4423.7	632.0	1.6424	0.1517
Residuals	40	15391.4	384.8		

Both t-tests estimate the `Treatment` effect at 16.8 ms.

Both the analysis of variance using quasi-F ratios as well as the by-subject and by-item analyses are rather unsatisfactory from the point of view of statistical modeling. The quasi-F test presupposes a model that requires 96 parameters to estimate 96 data points. A simple list would provide a more parsimoneous account of the data. Moreover, since the model overfits the data in the extreme, prediction is impossible and neither confidence intervals nor prediction intervals can be estimated for the coefficients. The conventional procedure of carrying out separate subject and item analyses, by contrast, comes with the problem that two models instead of one model are fit. Neither model provides accurate predictions, not even for the original data points. The by-subject analysis fails for the individual items, the by-item analysis fails for the individual subjects.

As in the preceding examples, multilevel modeling offers an alternative that is parsimoneous in the number of parameters, and that avoids overfitting the data. Again, we take `Subject` to be the main grouping factor for the random effects structure, and `Treatment` to be a fixed effect. Because multi-level modeling does not provide the option of crossing `Pair` as a random effect with `Subject`, we will consider a multilevel model in which we include `Pair` as a fixed effect while at the same time nesting `Pair` as a random effect under `Subject`.

Table 11 presents the least squares decomposition for this multilevel model fit to the data of Table 9, using an ordinary least-squares decomposition of the data. The Mean Squares for the error stratum `Subject`, for instance, 639.5, is identical to that listed in Table 10 for `Subject`. There are three fixed effects in this model: the fixed effect of `Treatment`, the fixed effect of `Pair`, and their interaction. There are also three random effects in this model: `Subject` (mean squares 639.5), `Pair` by `Subject` (mean squares 563.8, again as in Table 10), and the pooled error of the interaction of `Subject` by `Treatment` and the third order interaction. This pooled error is the error term for testing the main effect of `Treatment`.

There are two things to note at this point. First, unlike in the standard decomposition of the data shown in Table 10, we now have a non-zero error term. Hence, we have a model that will allow prediction to new data. Second, with this error term, we test whether the `Treatment` effect has explanatory value compared to both how subjects respond to the treatment and to how the treatment effect changes for combinations of pairs and subjects. Especially for experiments in which the treatment effect is implemented in the pairs — in the present example as a linguistic change in the form of one sentence of a pair resulting in the second sentence of the pair — this is a sensible choice, as an effect of the treatment on the subjects independently of the linguistic form in which this treatment is administered is uncontrolled random variation just as the third order interaction. Note that this way of testing for a `Treatment` effect differs from, e.g, the quasi-F test.

Crucially, multilevel modeling allows the researcher to go beyond the analysis of variance decomposition of Table 11, in that it provides estimates of all the parameters of the model, not only the parameters for the fixed effects, but also the parameters of the random effects. For the present example, these parameters are listed in Table 12 together with their 95% confidence intervals. An analysis of variance table (Table 13) shows very similar F-values and p-values as Table 11, even though these estimates are arrived at computationally in a completely different way, namely, with relativized maximum likelihood estimation instead of by ordinary least squares estimation.

Table 12 shows that we need 19 parameters to fit 96 data points, which is a considerable improvement over the 96 parameters required for carrying out a quasi-F test. This table also highlights that we have estimates for the sample effect of the sentence pairs. As in the preceding example, we can check whether the coefficients of `Pair` are normally distributed. Since a Shapiro-Wilk test of normality does not suggest any departure from normality (p = 0.6355), we can center the coefficients of `Pair` while at the same time adding the mean of these coefficients to the intercept. The new intercept, 211.18, is appropriate for predicting fixation durations for new, unseen items. Similarly, we can inspect the normality of the `Treatment` by `Pair` coefficients (p = 0.2996 according to the Shapiro-Wilk test), and center these coefficients, while simultaneously adding their mean to the `Treatment` effect, changing it from -20.576 to -16.777, in order to make prediction to unseen items possible.

Table 12
Estimates and 95% confidence intervals for the parameters of the random and fixed effects in the model fit to the data of Table 9.

	lower	estimate	upper
Fixed effects			
Intercept	170.757	188.816	206.875
`Treatment` (simple: complex)	-43.467	-20.576	2.314
`Pair`:			
contrast pair 1, pair 2	-15.452	10.070	35.593
contrast pair 1, pair 3	-23.361	2.161	27.684
contrast pair 1, pair 4	4.871	30.395	55.918
contrast pair 1, pair 5	15.578	41.102	66.625
contrast pair 1, pair 6	24.163	49.687	75.210
contrast pair 1, pair 7	-9.583	15.940	41.463
contrast pair 1, pair 8	4.068	29.592	55.115
`Treatment` by `Pair`			
complex, contrast pair 1, pair 2	-32.010	0.361	32.733
complex, contrast pair 1, pair 3	-15.918	16.453	48.826
complex, contrast pair 1, pair 4	-45.441	-13.068	19.303
complex, contrast pair 1, pair 5	-31.680	0.692	33.064
complex, contrast pair 1, pair 6	-48.503	-16.131	16.241
complex, contrast pair 1, pair 7	-10.356	22.015	54.387
complex, contrast pair 1, pair 8	-12.296	20.076	52.448
Random effects			
`Subject`	0.010	2.199	453.755
`Pair` within `Subject`	3.937	9.453	22.695
Residual	15.756	19.617	24.424

Since the interaction of `Treatment` by `Pair` is not significant, we might consider an alternative model in which we remove `Pair` altogether from the fixed effects structure, while retaining it as a random effect nested under `Subject`. This nesting captures the main effect of `Pair` implicitly. This second, less restrictive model turns out to be slightly more conservative with respect to the significance of the `Treatment` effect: $F(1, 47) = 16.018, p = 0.0002$. The estimated parameters of this second multilevel model and their 95% confidence intervals are listed in Table 14. For this second model, $r^2 = 0.682$, which is

Table 13
Anova table for the fixed effects of the multilevel model fitted to the data of Table 9.

	F			*p*
Intercept	F(1,40)	=	6159.661	< 0.0001
Treatment	F(1,40)	=	17.553	0.0001
Pairs	F(7,35)	=	5.529	0.0002
Treatment * Pair	F(7,40)	=	1.642	0.1517

Table 14
Estimated parameters and their 95% confidence intervals for a multilevel model fit to the data of Table 9 with Treatment as only fixed effect.

	lower	estimate	upper
Fixed effects			
Intercept	203.569	211.184	218.800
Treatment (simplex: complex)	-25.209	-16.776	-8.343
Random effects			
Subject	1.023621e-09	0.378	140169753
Pair within Subject	10.941	16.277	24.214
Residual	16.775	20.535	25.137

slightly higher than that of the previous model, for which $r^2 = 0.649$. Note that this is achieved with just 5 instead of 19 parameters.

Inspection of the confidence intervals in Tables 12 and 14 reveals two problems, however. First, the more restricted model in which Pair is included as a fixed effect has a coefficient for Treatment with a 95% confidence interval including zero. The probability that this coefficient might not be zero is only 0.0768. This is an indication that the Treatment effect might not be fully reliable.

Second, note that the estimate of the standard deviation is not well-bounded for the for the Subject random effect. For the multilevel model incorporating Pair as a fixed effect, the 95% confidence interval for Subject is $[0.010, 453.755]$, for a standard deviation estimated at 2.199. In other words, the confidence interval is two orders of magnitude larger than the estimate itself. This is an indication that there might be a problem with the assumptions underlying the model. For the model with Treatment as only fixed effect, the huge confidence interval (ranging from practically zero to 140 million) indicates that removing the fixed effect of Pair from the model has as its consequence that the Subject effect can no longer be properly estimated. This situation contrasts markedly with the model fit to the data on tone neutralization in Taiwanese, for which the 95% confidence intervals for the estimated parameters in Table 5 were all properly bounded.

For this example, the quasi-F test as well as the by-item and by-subject tests lead one to reject the null hypothesis that there is no `Treatment` effect. The two multilevel models that we have considered support this conclusion, but these models come with the warning that there might be a problem with the fit. In fact, this warning is justified, as the dataset of Table 9 was generated by a simulation model without a `Treatment` effect.

To illustrate the Type I error rate as well as the power for the standard statistical tests and the multilevel model when `Pair` is included as a fixed effect crossed with `Treatment`, and also as a random effect nested under `Subject`, consider Table 15. (As discussed above, the assumption that `Pair` is (also) a fixed effect can be relaxed when its coefficients turn out to be normally distributed.) The details of this simulation can be found in the appendix. In the simulated data sets, `Subject` and `Pair` were actually implemented as crossed random effects. Alternatively, `Pair` could have been nested under `Subject`. However, since the current standards proceed from the assumption that the two random effects are crossed, this is the way the simulated data were constructed. This makes it possible to illustrate the advantages and disadvantages of multilevel modeling for data sets where it only approximates the true structure of the data.

The proportions listed in Table 15 are based on 100 simulation runs. The column labeled F1, F2 lists the results for the common practice of requiring both F1 and F2 to be significant. The first section of this table summarizes the results for the case in which the simulation model actually contains both a `Treatment` effect and all three pairwise interactions involving the random effects `Pair` and `Subject`. The second section reports the results when there are no interactions in the model. Note that for this simulation, there is little advantage of carrying out separate F1 and F2 tests compared to the quasi-F test.

The third and fourth sections of Table 15 illustrate performance when there is no main effect of `Treatment` in the simulation model. The Type I error rate is nominal for all models when there is no `Treatment` effect and no interactions, but the Type I error rate is far above the nominal levels for the multilevel and to some extent for the by-subject analyses when there are interactions.

What this simulation illustrates is that the multilevel model is an excellent choice when there are no interactions involving the random effects. However, the high Type I error rate for the multilevel model when there are interactions illustrates that the present multilevel model with `Pair` nested under `Subject` will tend to lead to the incorrect conclusion of a main effect of `Treatment` when there is no such main effect in the simulation, in which `Subject` and `Pair` were

Table 15
Illustration of power and type I error rate for 5 models testing for an effect of `Treatment` (100 simulation runs).

With `Treatment`, with interactions					
alpha	multilevel	subj	item	quasiF	F1, F2
0.05	0.79	0.55	0.44	0.35	0.39
0.01	0.66	0.30	0.13	0.06	0.10
With `Treatment`, without interactions					
alpha	multilevel	subj	item	quasi-F	F1, F2
0.05	0.95	0.81	0.89	0.73	0.75
0.01	0.82	0.52	0.65	0.43	0.40
Without `Treatment`, with interactions					
alpha	multilevel	subj	item	quasi-F	F1, F2
0.05	0.28	0.19	0.05	0.04	0.04
0.01	0.15	0.07	0.01	0.00	0.00
Without `Treatment`, without interactions					
alpha	multilevel	subj	item	quasi-F	F1, F2
0.05	0.02	0.08	0.01	0.03	0.01
0.01	0.00	0.03	0.00	0.00	0.00

crossed rather than nested. In part, this is due to the fact that the multilevel model only approximates the actual structure of the simulated data, testing the `Treatment` effect agains a different (pooled) error term. But it is also due to the statistical ambiguity of data sets with many interactions. The multilevel model shows that in many cases, a data set that was in fact generated without a `Treatment` effect and with subjects and pairs crossed, could just as well have been generated by a simulation model with a `Treatment` effect and with `Pair` nested under `Subject`. In other words, data sets with the present complexity in terms of interactions will often be statistically ambiguous, in the sense that different models might equally well have generated the data. In the present example, we know what the generating model underlying a given data set is. In practice, this information is not available. Given the presence of interactions, the researcher's confidence in the multilevel model will depend on what empirical justification exists for either crossing `Pair` with `Subject` or alternatively nesting `Pair` under `Subject`.

For this design, one might adopt a conservative strategy and follow the quasi-F test or, equivalently, the combined by-subject and by-item analyses. Although for the present example, this strategy protects against erroneously concluding that there is a main effect of Treatment in 14 cases (for $\alpha = 0.01$), it comes with the price of erroneously concluding that there is no such effect in 42 to 56 cases (again for $\alpha = 0.01$) when in fact the effect is there. For the present example, this conservative strategy is not particularly useful.

For cases in which the quasi-F test (or the combined F1 and F2 tests) are more conservative than the multilevel model, it is often illuminating to inspect the data graphically by means of a trellis plot. Figure 8 plots the effect of Treatment for each combination of Subject and Pair for a simulation run without a Treatment effect in which the quasi-F test correctly reports a non-significant effect of Treatment, whereas the multilevel model suggests the presence of a Treatment effect. What this trellis plot shows is that there is considerable variation as to the direction of the Treatment effect. For 29 subject-pair combinations, the effect in the simple condition is smaller than in the complex condition. In 19 cases, the effect goes in the opposite direction. Even though the multilevel model suggests there might be a main effect of Treatment, the variation across subjects and sentence pairs severely restricts the interpretation of this main effect. There is only a small majority for the effect of Treatment leading to longer fixation durations in the complex condition, and given the substantial number of subject-item pairs in which there is a large effect in the opposite direction, this main effect is not informative at all. Thus, the disagreement between the quasi-F test, which argues against a main effect, and the multilevel model, which argues in favor of a main effect, concerns the question of whether there is a — potentially quite small — majority showing the effect in a specific direction. The answer to this question is co-determined by the reseacher's assumptions about whether or not the strong assumption of crossing Pair with Subject is justified.

In fact, reporting whether or not the main effect of Treatment is significant for this design is not informative and potentially misleading without additional information about the interactions in the data. It is a misunderstanding to assume that a main effect would allow the conclusion that the effect generalizes to 'the population'. Such a conclusion is only justified in case there are no demonstrable interactions. In the presence of interactions, the combined effect of treatment and its interactions should be considered. When there are strong interactions, as in Figure 8, the question whether there is a main effect of Treatment becomes a side issue.

Although being able to report a significant main effect seems to be widely regarded as a necessary condition for an experiment to have been successful, it should be kept in mind that an experiment without a main effect of Treatment but with clear interactions is not necessarily a failure. Imagine, for instance, that a pattern similar to that shown Figure 8 is observed in medical research on a treatment involving a new drug, with a minority of subjects showing an effect of the drug in the desired direction, and a majority showing an adverse

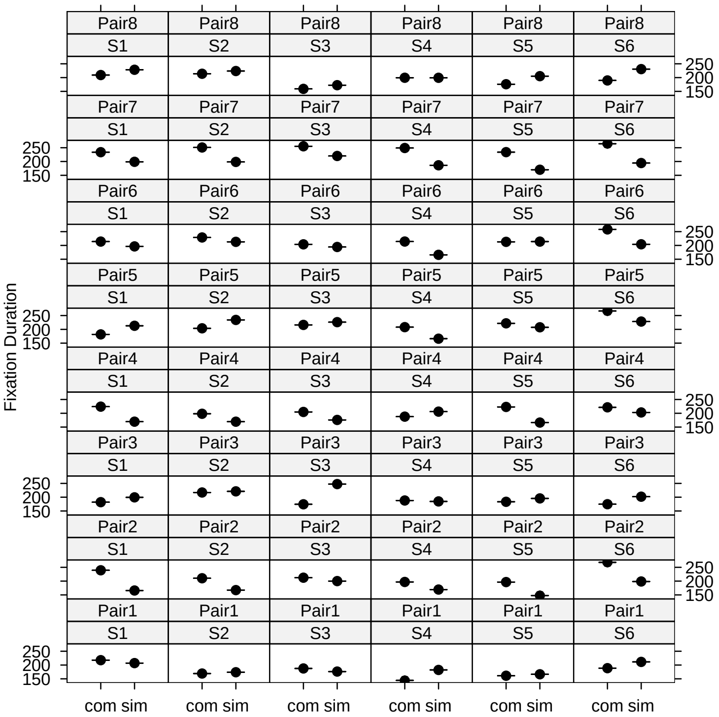

Figure 8. Fixation duration as a function of treatment for a simulated data set where the multilevel model reports a main effect of treatment while the quasi-F test reports no effect.

drug reaction. In such a situation, follow-up research is clearly required as to why only some subjects benefit from the new drug. Similarly, differences in the degree and direction of linguistic variables should be explored and understood, rather than ignored.

The emphasis in the psycholinguistic literature on the quasi-F test (Clark, 1972, and Raaijmakers et al., 1999) and the simultaneous lack of interest in how the interpretation of potential main effects depends on the interactions is, from the perspective of modern exploratory data analysis, both naive and wasteful.

4 Analyzing dichotomous variables

Experiments in which different kinds of responses are elicited, such as primed and unprimed lexical decision, number decision, and grammaticality decision, yield two kinds of dependent variables: latencies on the one hand, and on the other hand a binary decision variable with values such as correct/incorrect, singular/plural, and grammatical/ungrammatical. Most studies report analyses of such binary data using standard least squares analysis of variance and regression applied to proportions (e.g., proportions of errors by subject and by item). However, it is important to choose

> ...a model whose mathematical form is appropriate for the response being modeled. This often has to do with minimizing the need for interaction terms that are included only to address a basic lack of fit. For example, many researchers have used ordinary least squares regression models for binary responses, because of their simplicity. But such models allow predicted probabilities to be outside the interval $[0, 1]$, and strange interactions among the predictor variables are needed to make predictions remain in the legal range. (Harrell, 2001:7)

A second reason that the ordinary least squares method is unsuitable is that since Y is a binomial variable, its mean and variance are both linear in the probability of Y. In other words, Y is intrinsically heteroskedastic, violating one of the basic assumptions of ordinary least squares modeling.

The statistical model that is generally recommended for analyzing binary response data is the logistic model. If we code binary outcomes Y as 0 (failure) and 1 (success), a regression or analysis of variance model can be stated in terms of the probability that $Y = 1$ given the predictors $X_1, X_2, \ldots$, defined as follows:

$$\Pr(Y = 1|X_1, X_2, \ldots) = P = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2\ldots)}}. \tag{6}$$

This is equivalent to modeling the log odds ratio as a linear function of the predictors:

$$\begin{aligned}\text{logit}(Y = 1|X_1, X_2, \ldots) &= \text{logit}P = \log(\frac{P}{1-P}) = \\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 \ldots.\end{aligned}$$

The parameters of the model are estimated by the method of maximum likelihood, with significance being evaluated with Z-scores instead of t-scores. Logistic models are available in many statistical packages, and there is no reason other than unjustified conservatism and methodological laziness not to use them. Chatterjee et al. (2000), chapter 12, provides a very readable introduction to logistic modeling, Harrell (2001) provides extensive examples of more complicated data sets, including bootstrap validation and non-linear regression.

As an illustration of the advantage of logistic modeling, consider an experiment in which the dependent variable is the accuracy measure (correct versus incorrect response in, for instance, lexical decision). Let's assume that accuracy

Table 16
Comparison of models with the error proportions as dependent variable with logistic regression. Proportions are based on 1000 simulation runs.

α	model	`Treatment`		Covariate	
		proportion	logit	proportion	logit
0.05	with `Treatment`	0.491	0.501	0.909	0.922
0.01	with `Treatment`	0.252	0.258	0.763	0.805
0.05	without `Treatment`	0.088	0.046	0.853	0.868
0.01	without `Treatment`	0.015	0.008	0.680	0.712

is hypothesized to be a linear function of a binary treatment effect (singular versus plural number) and a discrete covariate (frequency of occurrence). Traditionally, such data are reported by calculating proportions of errors over subjects and items. In what follows, I will first compare a by-item analysis of the error proportions with a logistic analysis of the corresponding log odds ratios. After this, an example is provided of a single logistic analysis replacing the standard by-subject and by-item analyses.

Table 16 lists the proportions of simulation runs for which the treatment and frequency effects were reported as significant at the 0.05 and 0.01 significance levels, for analyses with the error proportion as the dependent variable, as well as for logistic analyses. All simulation models included an effect of the frequency covariate. As shown by the last two columns of Table 16, the logistic regression is characterized by slightly superior power. As the differences are small, in order to obtain reasonably stable estimates of the magnitude of these differences, the proportions in Table 16 were based on 1000 rather than 100 simulation runs. (Note that this is still an example only; for solid estimates, the number of simulation runs should be at least an order of magnitude larger.) When the model contained a `Treatment` effect, the logistic analyses outperformed the analysis using proportions by a tiny margin. However, when there was no effect of `Treatment` in the model, the analysis using proportions emerged with a slightly too high Type I error rate, whereas the logistic regression performed as required. For the details of this simulation, the reader is referred to the appendix.

Table 16 provides an example of the advantage of logistic regression in terms of power and Type I error rate. The final question to be addressed in this section is whether it is necessary to run separate by-item and by-subject analyses. Consider a data set with the accuracy measure as dependent variable, and in addition to the `Treatment` (singular versus plural) and `Frequency` effects of the preceding example, an additional longitudinal effect of fatigue. We can fit a logistic regression model to such a data set, where we use the `Frequency` covariate to bring item variability under control, and where we include `Subject`

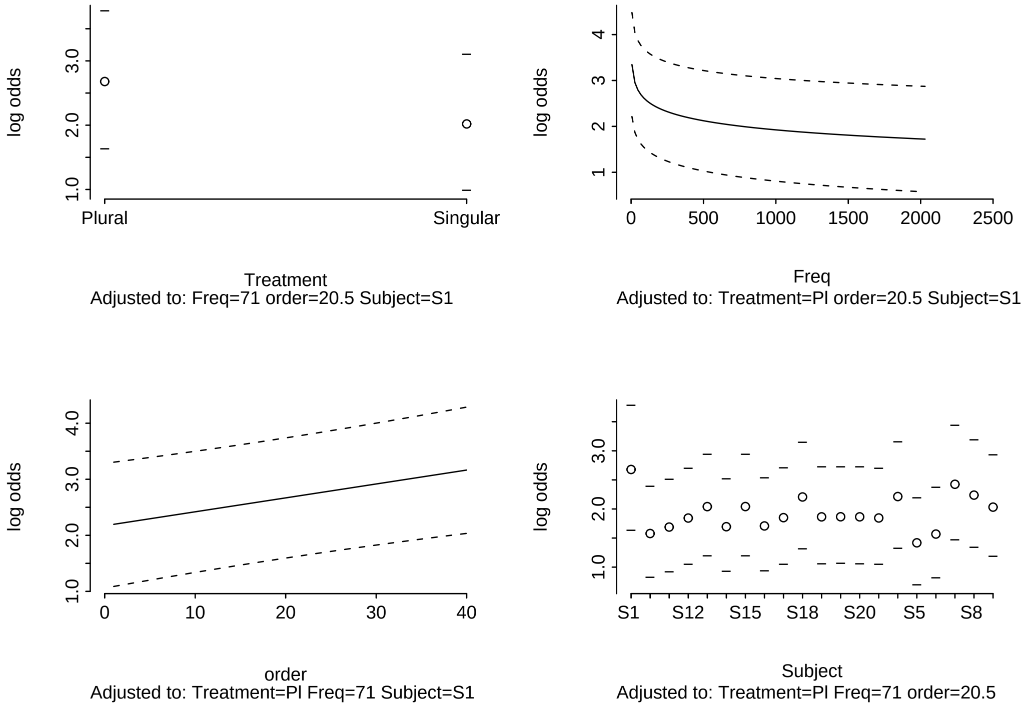

Figure 9. Log odds as a function of `Treatment` (upper left), `Frequency` (upper right), position in the sequence of items (`order`) (lower left), and `Subject` (lower right), with 95% confidence intervals.

as a (fixed) factor to bring subject variability explicitly into the model. (For ordinal logistic regression, see Harrell, 2001, and Sheu, 2002). If the effect of `Subject` is significant, and if the coefficients of the `Subject` effect are normally distributed, then the model can be reformulated with `Subject` as random effect by means of centering and adjustment of the intercept. Figure 9 illustrates the estimated marginal effects of the predictors on the log odds ratio, with 95% confidence intervals. Thus, instead of collapsing the data over subjects or items, we can fit the model directly to the individual data points (using the Poisson canonical link function instead of the logit, see, e.g., Venables & Ripley, 1994:185–199), with subject variability directly under control, and with item variability under control through the frequency covariate.

5 Concluding remarks

The main targets of the present critique of the current gold standards in psycholinguistics can be summarized as (1) the dictatorship of the factorial design, (2) the hegemony of prior averaging, and (3) unjustified methodological conservatism.

Factorial designs are commonly used where regression is more appropriate. Dichotomization and factorization of numerical predictors, although widely practised, lead to a loss of power and should be avoided. Psycholinguists are generally very reluctant to include covariates in their analyses, even though including relevant covariates is part and parcel of statistical common sense. When relevant covariates are not taken into account, the conclusions suggested by one's model may be unwarranted. Including available covariates is a crucial part of the modeling process, just as checking for outliers, handling of collinearity (see, e.g., Baayen, Feldman, & Schreuder, 2004), and addressing potential non-linearities (see, e.g., Harrell, 2001:230).

The examples discussed in the present study provide ample illustration of the disadvantages of the averaging procedures underlying the by-subject and by-item analyses. Prior averaging not only leads to a loss of power, but also makes it impossible to bring longitudinal effects in the experiment in a principled way into the model. Moreover, accurate prediction to the individual data points is impossible on the basis of the subject and item analyses. The quasi-F test advocated by Clark (1973) and Raaijmakers et al. (1999) for one kind of design is a not particularly useful alternative, due to its lack of power and the impossibility of insightful prediction. Often, multilevel modeling will provide a better tool for understanding the structure of the data.

Multilevel models are an excellent tool for analysing data with nested random effects. For repeated measures designs with items and subjects, this raises the question of whether items should be crossed with subjects or nested under subjects. There is no reflection in the literature (with the exception of Quené & van den Berg, 2001) about the advantages and disadvantages of crossing items with subjects in combination with an item by subject interaction, compared to nesting items under subjects. The traditional approach, guided by practical expedience, is to proceed under the assumption that subjects and items are crossed. Modeling items as crossed with subjects entails a strong assumption about the commonality of the specific item effects across subjects that is absent when modeling items as nested under subjects. In the multilevel approach, one possibility is to nest items under subjects, in which case common item effects across subjects, if present, are modeled implicitly. Alternatively, item can initially be included as a fixed effect and at the same time as a random effect nested under subject. If the coefficients of the fixed effect of item turn out to be normally distributed, the model can be reformulated with item as a random effect, to allow prediction to novel, unseen items. Of course, prediction is meaningful only when the items in the experiment are a true random sample from the population of items. When items have been carefully screened for various properties, items should be analyzed as fixed, or the characteristics of the sub-population for which prediction is envisioned should be made explicit.

From a methodological perspective, it is clear that in psycholinguistics, the discussion initiated by Clark (1973) has led to a situation in which the techniques available in the 1970s for analysis of variance have become the gold standard and reign supreme, as if the current gold standard of by-subject and by-item analyses were the best that modern statistics has to offer. The importance ascribed to these analyses is illustrated by the following recommendation to its reviewers on the homepage of the *Journal of Memory and Language*:

> In general, it is important to be confident that the results generalize over items as well as over participants.

Classical statistics seems to be viewed as perfected and finished, with the by-subject and by-item analyses as the optimal solution for repeated measurement problems that are sometimes naively thought to be specific to the field of psycholinguistics. Multilevel modeling and logistic modeling are shrugged away as theoretical variations on a well-established theme with no practical benefit for the data analyst. Likewise, the idea that trellis graphs might enhance insight over and above the numbers produced by standard statistical packages sometimes meets surprising scepticism. However, as anyone following statistical developments outside the specific field of psycholinguistics (for instance, in *Psychological Methods* or in *Behavioral Research Methods, Instruments and Computers*, or in Venables & Ripley, 2003) will have realized, current statistics has a lot more to offer, both in power and in the insight provided into the quantitative structure of the data.

Unfortunately, those researchers who do read up on the literature or who enlist the help of professional statisticians to analyse their data are often forced by the review process to ostracise the resulting more sophisticated data analyses from their paper, substituting it for the traditional by-subject and by-item analyses. Whereas reviewers should read up on statistical techniques unknown to them, they tend to misuse their authority to bring the statistical analyses back in line with the current gold standard. In fact, many of my colleagues admit knowing that they use non-optimal statistical techniques, but regard getting the data published as more important as getting the data published with the correct data analyses. This attitude has led to an extreme form of unjustified methodological conservatism.

In this respect, the preface to the 14th edition of Yule and Kendall's *Introduction to the theory of statistics* from 1950 is instructive. Kendall explains that

> Although fewer than fifteen years have passed since the last revision, so much has happened in the statistical world in the meantime that Mr. Yule and I both felt that the usefulness of the book would be increased by some further changes. (page v)

In the same vein, a lot has happened in statistics since the 1970s (logistic models, multilevel models, classification and regression trees, regression splines, trellis graphs, bootstrap validation), and psycholinguistic research would benefit if the journals would allow researchers the freedom to make use of the new developments, rather than impeding research by imposing standards of the past.

Appendix

All simulations were run using the `R` statistical programming environment (`http://lib.stat.cmu.edu/R/CRAN/`) version 1.7.0 and the `nlme` library of Pinheiro and Bates (2000), version 3.1-45. Figure 9 was created with the `Design` library of Harrell (2001), version 2.0-2.

Simulation example 1 (Figure 1, Table 1): *A comparison of an analysis of variance with a regression analysis.* These two methods were compared for the following models:

$$Y1 = -60 + 3.8X + \epsilon,\ \epsilon \sim N(0, 80^2),\ X \sim \text{Unif}(-10, 6.6) \tag{7}$$

$$Y2 = -X^2 + \epsilon,\ \epsilon \sim N(0, 80^2),\ X \sim \text{Unif}(-10, 6.6) \tag{8}$$

$$Y3 = 500 - 4\log(X) + \epsilon,\ \epsilon \sim N(0, 20^2),\ X \sim \text{Unif}(1, 1000) \tag{9}$$

For each of 100 simulation runs with 1000 random values of X, uniformly distributed on the abovementioned intervals, 15 data points were randomly sampled from the 5% lowest and 5% highest values of X for the factorial contrast, and 30 data points were randomly sampled from the full data range for the regression analyses. For illustration purposes, the standard deviations for the data sets shown in the bottom panels of Figure 1 were set to 10 instead of 20.

Simulation example 2 (Figure 2, Table 2): *A comparison of analysis of variance with regression for two continuous correlated predictors.* In order to construct these correlated predictors, let $T_1 \sim \text{N}(3, 6^2)$, $T_2 \sim \text{N}(6, 6^2)$, and $C \sim \text{N}(25, 5^2)$, and define

$$X = T_1 + C,$$
$$Y = T_2 + C.$$

We consider two models, one with and one without an effect of X:

$$\text{RT}_1 = 400 + 3X + 6Y + \epsilon, \tag{10}$$

$$\text{RT}_2 = 400 + 0X + 6Y + \epsilon, \epsilon \sim \text{N}(0, 50^2). \tag{11}$$

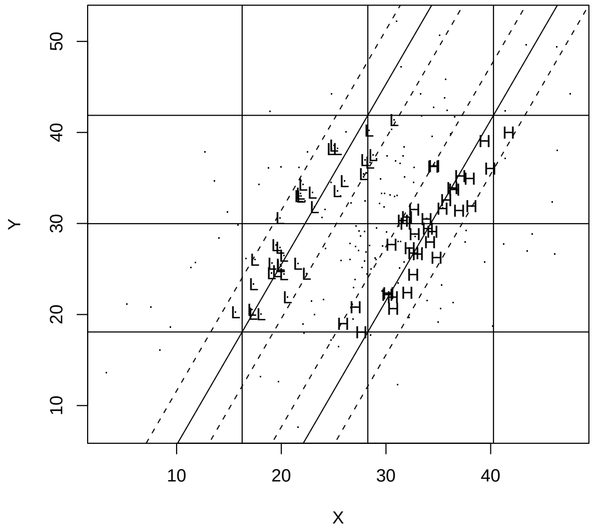

Figure 10. The geometry of item selection for simulation 2.

Figure 10 illustrates how for each simulation run a factorial contrast in X matching in the mean for Y was constructed. The central vertical line represents the sample mean $\bar{X}$, the vertical lines to its left and right are at a distance of $1.5\hat{\sigma}_X$ of $\bar{X}$. The central horizontal line is at the sample mean of Y, the upper and lower horizontal lines are at a distance of $1.5\hat{\sigma}_Y$ of $\bar{Y}$. The dashed lines are at a distance of $0.375\hat{\sigma}_X$ of the solid slanted lines. Items were selected from each of the central four sections enclosed by dashed lines and horizontal lines. This ensures that any two data points within a narrow interval of Y are at least $0.75\hat{\sigma}_X$ apart. A number equal to the minimum of the number of data points in any of these sections (but not exceeding 40) was randomly selected from each of the sections. For the regression analysis, a number of data points equal to the total number of data points in the analysis of variance was selected randomly from the complete sample.

Simulation example 3 (Table 4): *Power and Type I error for multiple regression comparing multilevel regression, item regression, and random regression.* The simulated data sets were generated by the models

$$\mathrm{RT}_{ij} = \overbrace{400 + 3X_i + 6Y_i + 4.5Z_i}^{\text{fixed effects}} + \overbrace{b_j}^{\text{random effects}} + \epsilon_{ij} \tag{12}$$

$$\mathrm{RT}_{ij} = \overbrace{400 + 3X_i + 6Y_i + 0Z_i}^{\text{fixed effects}} + \overbrace{b_j}^{\text{random effects}} + \epsilon_{ij} \tag{13}$$

$$i = 1, 2, \ldots, 20;\ j = 1, 2, \ldots, 10,$$
$$b_j \sim \mathrm{N}(0, 4^2),\ \ \epsilon_{ij} \sim \mathrm{N}(0, 250^2),$$

for 20 items and 10 subjects. The predictor X always had the values 1, 2, $\ldots$, 20. The predictors Y and Z were (nearly) orthogonal and uniformly distributed on the interval $[1, 20]$. In this simulation, as in the simulations to follow, the random effects were random samples from a population with zero mean, without further constraints, even though the estimated random effects are always constrained to sum to zero in the models fitted to the simulated data. This choice is motivated by the consideration that in actual data, the true random effects that happen to appear in the sample will seldom sum up to zero.

Simulation example 4 (Table 6, Table 7, Figure 7): The analysis of a simple Latin Square design. The simulated data sets were generated by the model

$$\begin{aligned}
\mathrm{RT}_{ijk} = & \overbrace{534 + 4X_{t[ij]} + \mathrm{SOA}_k}^{\text{fixed effects}} + \overbrace{b_i + b_j}^{\text{random effects}} + \epsilon_{ijk}, \\
& i, j = 1, 2, \ldots, 12;\ X_{t[ij]} = 1, 2, \ldots, 12; \\
& b_i \sim \mathrm{N}(0, 28^2),\ b_j \sim \mathrm{N}(0, 39^2),\ \epsilon_{ijk} \sim \mathrm{N}(0, 20^2), \\
& \mathrm{SOA}_1 = 0(\text{long}),\ \mathrm{SOA}_2 = 5(\text{short}),\ \mathrm{SOA}_3 = -4(\text{medium}),
\end{aligned} \tag{14}$$

where i ranges over items, j ranges over subjects, and $t[ij]$ is the effect of fatigue at trial t for item i and subject j. The random effects of subject and item were uncorrelated. For the simulations without a `Treatment` effect, the coefficients for both dummy variables were set to zero. For the models without a learning effect, the coefficient of $X_{t[ij]}$ was set to zero. Multilevel modeling is especially recommended for this kind of incomplete design (see, e.g., Venables & Ripley, 2003:282). The multilevel model fit to this data set was

$$\begin{aligned}
\mathrm{RT}_{ijk} = & \overbrace{\beta_0 + \beta_1 X_{t[ij]} + \mathrm{SOA}_k + \mathrm{Item}_i}^{\text{fixed effects}} + \overbrace{b_j}^{\text{random effects}} + \epsilon_{ijk}, \\
& k = 1, 2, 3;\ i, j = 1, 2, \ldots, 12; \\
& b_j \sim \mathrm{N}(0, \sigma_b^2),\ \epsilon_{ijk} \sim \mathrm{N}(0, \sigma_\epsilon^2).
\end{aligned} \tag{15}$$

When the coefficients of `Item` are normally distributed, we can reformulate this fixed effect as a random effect, leading to the model

$$\begin{aligned}
\mathrm{RT}_{ijk} = & \overbrace{\beta_0' + \beta_1 X_{t[ij]} + \mathrm{SOA}_k}^{\text{fixed effects}} + \overbrace{b_j + b_i}^{\text{random effects}} + \epsilon_{ijk}, \\
& k = 1, 2, 3;\ i, j = 1, 2, \ldots, 12; \\
& b_j \sim \mathrm{N}(0, \sigma_{b_j}^2),\ b_i \sim \mathrm{N}(0, \sigma_{b_i}^2),\ \epsilon_{ijk} \sim \mathrm{N}(0, \sigma_\epsilon^2),
\end{aligned} \tag{16}$$

where β_0' is the sum of β_0 and the mean of the coefficients of `Item`, and where the b_i are the centered coefficients of `Item`.

Simulation example 5 (Table 15) *A split-plot design requiring a quasi-F test in classical analysis of variance.* Let D denote the dependent variable, let i and j index pairs and subjects, and let T denote `Treatment` ($T_1 = 15$ for the complex condition and $T_2 = 0$ in the simplex condition). The model generating the data sets is

$$\begin{aligned} D_{ijk} = & \overbrace{200 + T_k}^{\text{fixed effects}} + \overbrace{b_i + b_j + b_{ij} + b_{ik} + b_{jk}}^{\text{random effects}} + \epsilon_{ijk} \\ & i = 1, 2, \ldots, 8;\ j = 1, 2, \ldots, 6; k = 1, 2, \\ & b_i \sim \mathrm{N}(0, 7^2),\ b_j \sim \mathrm{N}(0, 7^2),\ b_{ik} \sim \mathrm{N}(0, 12^2),\ b_{jk} \sim \mathrm{N}(0, 4^2), \\ & b_{ij} \sim \mathrm{N}(0, 8^2),\ \epsilon_{ijk} \sim \mathrm{N}(0, 20^2). \end{aligned} \tag{17}$$

In the simulations without interactions, all b_{ij}, b_{ik}, b_{jk} were set to zero. In the simulations without a treatment effect, T_1 was set to zero. All random effects were pairwise uncorrelated. The multilevel model fit to the simulated data sets was

$$\begin{aligned} D_{ijk} = & \overbrace{\beta_0 + T_k + \mathrm{PAIR}_i + T \cdot \mathrm{PAIR}_{ik}}^{\text{fixed effects}} + \overbrace{b_j + b_{i(j)}}^{\text{random effects}} + \epsilon_{i(j)jk} \\ & b_j \sim \mathrm{N}(0, \sigma^2_{b_j}),\ b_{i(j)} \sim \mathrm{N}(0, \sigma^2_{b_{i(j)}}),\ \epsilon_{i(j)jk} \sim \mathrm{N}(0, \sigma^2_\epsilon). \end{aligned} \tag{18}$$

When centering the effects involving `Pair` is appropriate, this model can be reformulated as

$$\begin{aligned} D_{ijk} = & \overbrace{\beta'_0 + T'_k}^{\text{fixed effects}} + \overbrace{b_j + b_i + b_{ik} + b_{i(j)}}^{\text{random effects}} + \epsilon_{i(j)jk} \\ & b_j \sim \mathrm{N}(0, \sigma^2_{b_j}),\ b_i \sim \mathrm{N}(0, \sigma^2_{b_i}), \\ & b_{ik} \sim \mathrm{N}(0, \sigma^2_{b_{ik}}),\ b_{i(j)} \sim \mathrm{N}(0, \sigma^2_{b_{i(j)}}),\ \epsilon_{i(j)jk} \sim \mathrm{N}(0, \sigma^2_\epsilon), \end{aligned} \tag{19}$$

where β'_0 equals the sum of β_0 and the mean of the coefficients of `Pair`, and where T'_k is the sum of T_k and the mean of the coefficients of the `Pair` by `Treatment` interaction adjusting the k-th level of `Treatment`. The b_i are the centered coefficients of `Pair`, the b_{ik} are the centered coefficients of the `Pair` by `Treatment` interaction.

Simulation example 6 (Table 16) *Analysis of covariance using least squares estimation for proportions and maximum likelihood estimation for logits.* In this example, the dependent variable B was binary, with values 1 and 0. Let i and j index items and subjects, and let T denote the dummy variable for the `Treatment` contrast ($T_1 = 0, T_2 = 1$). The probabilities in this model were defined as

$$p(B_{ijk} = 1) = [1 + \exp\{-(-1 + 0.4T_k - 0.25F_i)\}]^{-1}, \quad i = 1, 2, \dots 40;\ j = 1, 2, \dots, 20;\ k = 1, 2, \tag{20}$$

with the 'frequency' F a compound lognormal-poisson random variable with mean 4 and standard deviation 1.5. For each subject j, the dependent variable B_{ijk} was set to 1 whenever a random number from the (0,1) interval was less than $p(B_{ijk} = 1)$, and to zero otherwise. In the analysis without an effect of `Treatment`, the coefficient 0.4 was replaced by 0.

References

Baayen, R., Feldman, L., Schreuder, R., 2004. A multivariate study of simple word recognition. Submitted.

Baayen, R. H., Dijkstra, T., Schreuder, R., 1997. Singulars and plurals in Dutch: Evidence for a parallel dual route model. Journal of Memory and Language 36, 94–117.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., 2003. Visual word recognition for single-syllable words. Submitted.

Balota, D. A., Chumbley, J. I., 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception and Performance 10, 340–357.

Becker, R. A., Cleveland, W. S., Shyu, M. J., 1996. The design and control of trellis display. Journal of Computational and Statistical Graphics 5, 123–155.

Becker, R. A., Cleveland, W. S., Shyu, M. J., Kaluzny, S., 1995. A tour of trellis display. Available: `http://cm.bell-labs.com/cm/ms/departments-/sia/wsc/webpapers.html`, Bell Labs.

Bertram, R., Schreuder, R., Baayen, R. H., 2000. The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity. Journal of Experimental Psychology: Learning, Memory, and Cognition 26, 419–511.

Bryk, A., Raudenbusch, S., 1992. Hierarchical linear models for social and behavioral research. Sage, Newbury Park, CA.

Clahsen, H., Eisenbeiss, S., Sonnenstuhl-Henning, I., 1997. Morphological structure and the processing of inflected words. Theoretical Linguistics 23, 201–249.

Clark, H., 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12, 335–359.

Cleveland, W. S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

Cobb, G. W., 1998. Introduction to design and analysis of experiments. Springer Verlag, Berlin.

Cochran, W. G., 1951. Testing a linear relation among variances. Biometrics 7, 17–32.

Cohen, J., 1983. The cost of dichotomization. Applied Psychological Measurement 7, 249–254.

Forster, K., 2000. The potential for experimenter bias effects in word recognition experiments. Memory & Cognition 28, 1109–1115.

Forster, K., Dickinson, R., 1976. More on the language-as-fixed effect: Monte-Carlo estimates of error rates for F_1, F_2, F', and *min*F'. Journal of Verbal Learning and Verbal Behavior 15, 135–142.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., Lafferty, T., 1987. The word frequency effect in lexical decision: Finding a frequency-based component. Memory and Cognition 15, 24–28.

Goldstein, H., 1995. Multilevel statistical models. Halstead Press, New York.

Haerdle, W., 1991. Smoothing Techniques With Implementation in S. Springer-Verlag, Berlin.

Harrell, F. E., 2001. Regression modeling strategies. Statistics and Computing. Springer, New York.

Hyönä, J., Pollatsek, A., 1998. Reading finnish compound words: Eye fixations are affected by component morphemes. Journal of Experimental Psychology: Human Perception and Performance 24, 1612–1627.

Jescheniak, J. D., Levelt, W. J. M., 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. Journal of Experimental Psychology: Learning, Memory and Cognition 20 (4), 824–843.

Levelt, W. J. M., Wheeldon, L., 1994. Do speakers have access to a mental syllabary. Cognition 50, 239–269.

Lindley, D., Smith, A., 1972. Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B 34, 1–42.

Lorch, R. F., Myers, J., 1990. Regression analyese of repeated measures data in cognitive research. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 149–157.

Meeuwissen, M., Roelofs, A., & Levelt, W. J. M., in press. Planning levels in naming and reading complex numerals. Memory & Cognition.

Myers, J., Tsay, J., 2002. Neutralization in taiwanese tone sandhi, unpublished manuscript, National Chung Cheng University.

Pinheiro, J. C., Bates, D. M., 2000. Mixed-effects models in S and S-PLUS. Statistics and Computing. Springer, New York.

Quené, H., Van den Bergh, H., 2001. On multi-level modeling as a remedy against the "language as fixed effect fallacy". Manuscript OTS, University of Utrecht.

Raaijmakers, J., Schrijnemakers, J., Gremmen, F., 1999. How to deal with "the language as fixed effect fallacy": common misconceptions and alternative solutions. Journal of Memory and Language 41, 416–426.

Satterthwaite, F. E., 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin 2, 110–114.

Sereno, J., Jongman, A., 1997. Processing of English inflectional morphology. Memory and Cognition 25, 425–437.

Sheu, C.-F., 2002. Fitting mixed effect models for repeated ordinal outcomes with the NLMIXED procedure. Behavioral Research Methods, Instruments and Computers 34, 151–157.

Taft, M., 1979. Recognition of affixed words and the word frequency effect. Memory and Cognition 7, 263–272.

Venables, W. N., Ripley, B. D., 2003. Modern Applied Statistics with S-Plus, 4th Edition. Springer, New York.

Wickens, T., Keppel, G., 1976. On the choice of design and of test statistic in the analysis of experiments with sampled materials. Journal of Verbal Learning and Verbal Behavior 22, 296–309.

Yule, G., Kendall, M., 1950. An introduction to the theory of statistics, 14th Edition. Charles Griffin & Company, London.