Brain and Language **81,** 55–65 (2002)
doi:10.1006/brln.2001.2506, available online at http://www.idealibrary.com on IDEAL®

The Subjects as a Simple Random Effect Fallacy: Subject Variability and Morphological Family Effects in the Mental Lexicon

R. Harald Baayen,* Fiona J. Tweedie,† and Robert Schreuder*

**Interfaculty Research Unit for Language and Speech, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; and †Department of Statistics, University of Glasgow, Glasgow, United Kingdom*

Published online November 28, 2001

This is a methodological study addressing the appropriateness of standard by-subject and by-item averaging procedures for the analysis of repeated-measures designs. By means of a reanalysis of published data (Schreuder & Baayen, 1997), using random regression models, we present a proof of existence of systematic variability between participants that is ignored in the standard psycholinguistic analytical procedures. By applying linear mixed effects modeling (Pinheiro & Bates, 2000), we call attention to the potential lack of power of the by-subject and by-item analyses, which in this case study fail to reveal the coexistence of a facilitatory family size effect and an inhibitory family frequency effect in visual and auditory lexical processing. © 2001 Elsevier Science (USA)

INTRODUCTION

This article addresses the vexed question of which statistical techniques are appropriate for the analysis of repeated-measures designs. A typical psycholinguistic experiment elicits response latencies from many different subjects to a large number of items. In a subject analysis, the response latencies of the participants are studied, averaged over the items. In an item analysis, the response latencies to the items are studied, averaged over the participants. The aim of this article is to show that the averaging processes underlying such analyses may be unwarranted due to systematic variability between subjects and that it may also mask the statistical relevance of experimental factors.

The materials reanalyzed in this study consist of 36 simplex words, varying on the dimensions of Surface Frequency (SurfFreq), Base Frequency (BaseFreq), Family Size (Vf), and Family Frequency (Nf); see Appendix C of Schreuder and Baayen (1997) for detailed information on the words used. The Surface Frequency of a word is the frequency of the word form itself. The Base Frequency of a word is the summed frequency of all inflectional variants of this word. The Family Size of a word is the type count of all complex words that contain this word as a morphological constituent, its morphological family members. Finally, the Family Frequency is the summed token frequency of these family members. The role of Surface Frequency and Base

Address correspondence and reprint requests to R. H. Baayen, Interfaculty Research Unit for Language and Speech, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH, Nijmegen, The Netherlands. E-mail: baayen@mpi.nl.

0093-934X/01 $35.00

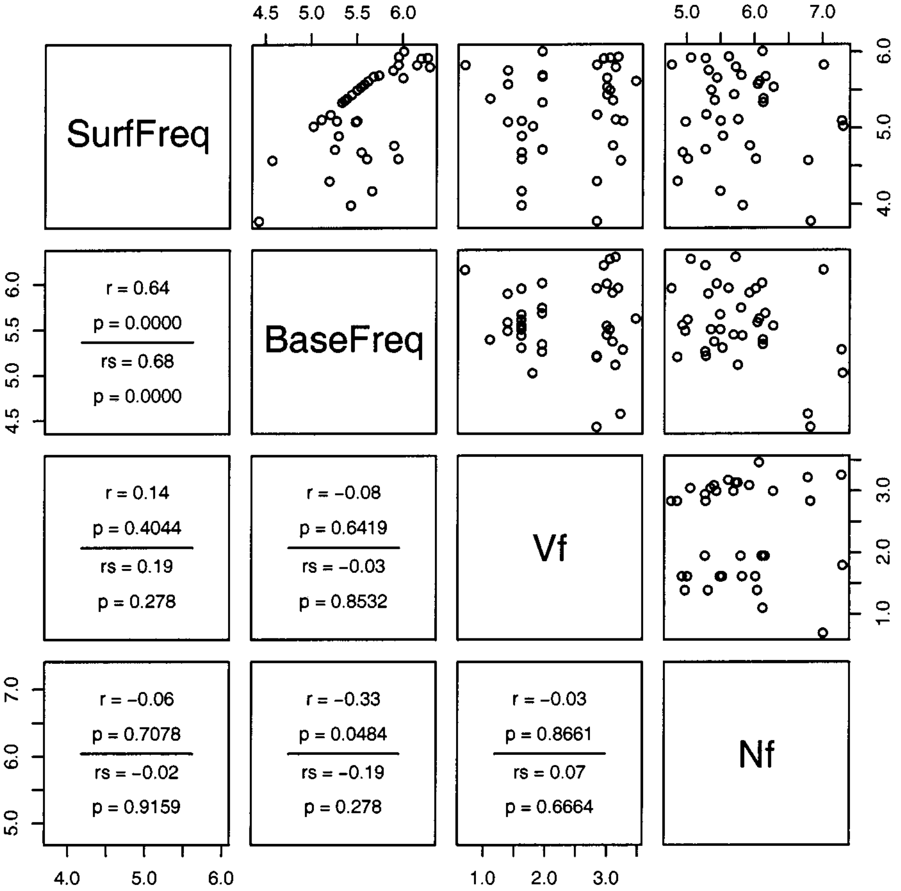

FIG. 1. Pairs plot for the properties of the experimental words. SurfFreq = log Surface Frequency; BaseFreq = log Base Frequency; Vf = log Family Size; Nf = log Family Frequency. Points represent words. The lower left triangle lists the corresponding Pearson (r) and Spearman (r_s) correlations.

Frequency is documented in, for instance, Taft (1979) and in Bertram, Schreuder, and Baayen (2000). Words with a higher Surface or Base Frequency are recognized more quickly than words with a lower Surface or Base Frequency. The facilitatory effect in visual lexical decision and subjective frequency rating of a large Family Size compared to a low Family Size is reported in Schreuder and Baayen (1997); Bertram, Baayen, and Schreuder (2000); and De Jong, Schreuder, and Baayen (2000). These studies report that the Family Frequency does not predict response latencies.

Figure 1 is a pairs plot of the correlation structure of the four independent variables in our design. For each pair of frequency measures, a scattergram plots the words in the space spanned by the logarithmically transformed frequency values. We transformed the frequency counts logarithmically for two reasons; first, because this removes most of the skewness from the frequency distributions; and second, because frequencies are perceived logarithmically (Rubenstein & Pollack, 1963; Scarborough, Cortese, & Scarborough, 1977). Figure 1 shows that log Surface Frequency and log Base Frequency are highly correlated. Collinearity diagnostics (Belsley, Kuh, & Welsch, 1980) confirm the impression that there is collinearity in the data matrix. The collinearity index for this data set is quite high, 61.61, and arises not only at the level of pairwise correlations (visible in Fig. 1) but also at the level of triadic correlation structure between Surface Frequency, Base Frequency, and Family Fre-

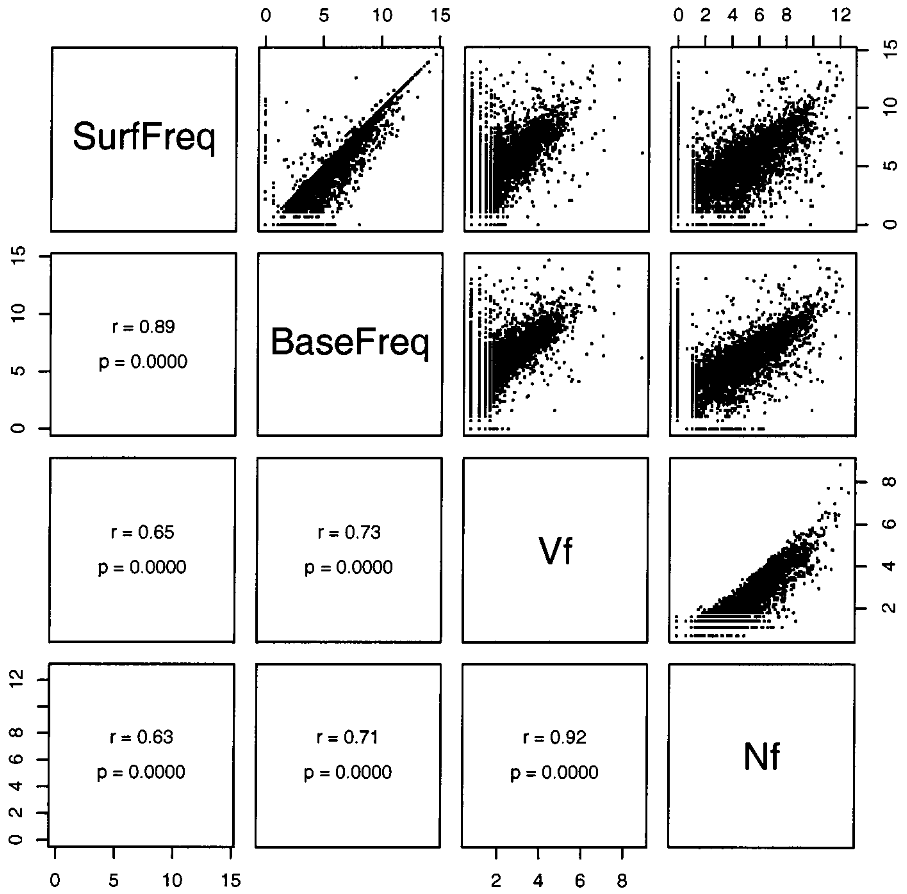

FIG. 2. Pairs plot for the distributional properties of monomorphemic words in Dutch as available in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). SurfFreq = log Surface Frequency; BaseFreq = log Base Frequency; Vf = log Family Size; Nf = log Family Frequency. Points represent words. The lower left triangle lists the corresponding Pearson correlations.

quency jointly (not visible in Fig. 1). This collinearity is a recurring problem in studies of these four frequency measures, which are all highly intercorrelated, as shown for Dutch monomorphemic words in Fig. 2 for the pairwise correlation structure.

This set of words was first used by Schreuder and Baayen (1997) in their Experiment 3. The original factorial design of this experiment, which contrasted a set of words with a large Family Size with a set of words with a low Family Size while controlling for the other frequency measures, is clearly visible in Fig. 1. The lexical decision latencies of 28 participants revealed a significant effect of Family Size, in both the by-item and the by-subject analyses. Response latencies for the same materials were also elicited from 39 participants using auditory lexical decision. This time, however, the factorial contrast was significant only in the by-subject analysis (see De Jong et al., 2000, p. 331, footnote 1).

In what follows, we show that the traditional by-subject and by-item analyses, due to the massive averaging processes involved, reveal only part of the structure in the data. We first discuss evidence that the assumption underlying the legitimacy of this averaging process, namely that individual participants do not differ systematically with respect to the independent variables, is incorrect for our experiments. We then proceed to show that a statistical technique that does not rely on these averaging processes reveals more independent variables to play a role in lexical processing than one would otherwise be led to believe. Notably, it will become clear that both Family Size and Family Frequency predict response latencies, with a high Family Size leading to shorter response latencies and a high Family Frequency leading to longer response latencies in both the visual and the auditory modalities.

PARTICIPANTS: STRUCTURED RANDOM EFFECTS

The averaging process underlying the by-item and by-subject analyses is perhaps defensible in case any variability in the response latencies of the individual participants is random with respect to fixed effect factors such as Surface Frequency, Base Frequency, Family Size, and Family Frequency. However, in the present data set, participants' responses are not random at all with respect to the fixed effect factors. This becomes apparent when we study the coefficients of linear models fitted to the responses of each individual participant, using the following random regression model:

$$E[y_{ij}] = \beta_{0i} + \beta_{1i}\mathrm{SurfFreq}_j + \beta_{2i}\mathrm{BaseFreq}_j + \beta_{3i}\mathrm{V}f_j + \beta_{4i}\mathrm{N}f_j, \qquad (1)$$

where y_{ij} is the log reaction time recorded for subject i for word j with associated log Surface Frequency, log Base Frequency, log Family Size, and log Family Frequency. Crucially, each subject i has their own fitted line to the words. In other words, given 28 participants, we inspect 28 separate linear regression models. The analysis is carried out on all correct responses, without any data cleaning—the error rate is very low—in order to avoid any possibility of data manipulation.

Figure 3 is a pairs plot of the coefficients for the Intercept, log Surface Frequency, log Base Frequency, log Family Size, and log Family Frequency for the visual lexical decision experiment. In a given panel in the upper right triangle of scatterplots, each dot represents a participant. The panels in the lower left triangle show the corresponding Pearson and Spearman correlations. We observe substantial correlations, notably between the Intercept and Base Frequency and the Intercept and Family Frequency as well as between Surface Frequency and Base Frequency. This suggests that there is systematic differential sensitivity among the participants with respect to Surface Frequency, Base Frequency, and Family Frequency.

Further statistical validation of this result is required, however. First, we need to center the four independent variables. Without centering, artificial correlations of the coefficients of a dependent variable with coefficients of the intercept may arise (Pinheiro & Bates, 2000, p. 34). For Surface Frequency, for instance, centering entails that we have to subtract the mean log Surface Frequency from each individual log Surface Frequency value.

Second, we need to make sure that the correlations that we observe between the sets of coefficients do not arise due to the substantial collinearity present in our data matrix. Hence, we use principal components regression (see Hocking, 1996, pp. 287–292). We orthogonalize the explanatory variables by transforming them to their principal components. The model is then fitted to the principal components, which are discarded based on the size of their associated eigenvalues. The final coefficients are then back-transformed to the original variables. This method allows us to deal with the greatly increased variability of coefficients that collinearity may cause, while keeping all of the variables in the model.

For these data, we discard the two principal components with the smallest associated eigenvalues. This removes 30.05% of the variation present in the data, allowing

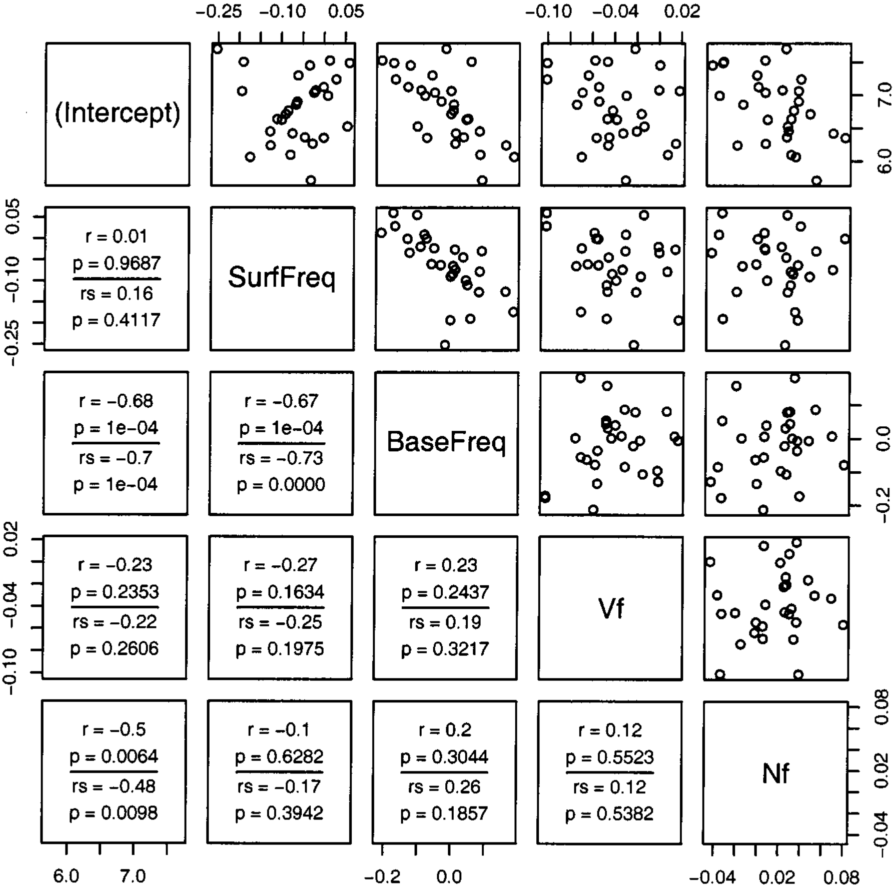

FIG. 3. Pairs plot for the coefficients of separate linear models fit to the responses of individual participants in visual lexical decision. SurFreq = log Surface Frequency; BaseFreq = log Base Frequency; Vf = log Family Size; Nf = log Family Frequency. No centering; points represent individual participants.

us to concentrate our attention on the most important 69.5% with a conservative model with only two principal components. The resulting coefficients for each subject are shown in Fig. 4. Inspection of the residuals of the models indicates that the models fit the data reasonably well. First note that there are no significant correlations of the independent variables with the intercept. We conclude that the correlations with the intercept visible in Fig. 3 are indeed an artifact of noncentered data analysis.

For the coefficients of the independent variables themselves, it is clear that there is an interesting correlation structure. Surface Frequency and Base Frequency are positively correlated: The more sensitive a subject is to Surface Frequency, as indicated by a more negative coefficient for Surface Frequency, the more sensitive this subject will be to Base Frequency, likewise indicated by a more negative coefficient for Base Frequency. (Note that this positive correlation differs from the negative correlation observed in Fig. 3. Apparently, this negative correlation is an artifact of the strong collinearity of Surface Frequency and Base Frequency.)

A second positive correlation emerges for Surface Frequency and Family Size, a correlation that was not visible in the analysis based on the raw data. Participants for whom Surface Frequency has a relatively large facilitatory effect are also participants for whom Family Size has a large facilitatory effect. Since Family Size taps

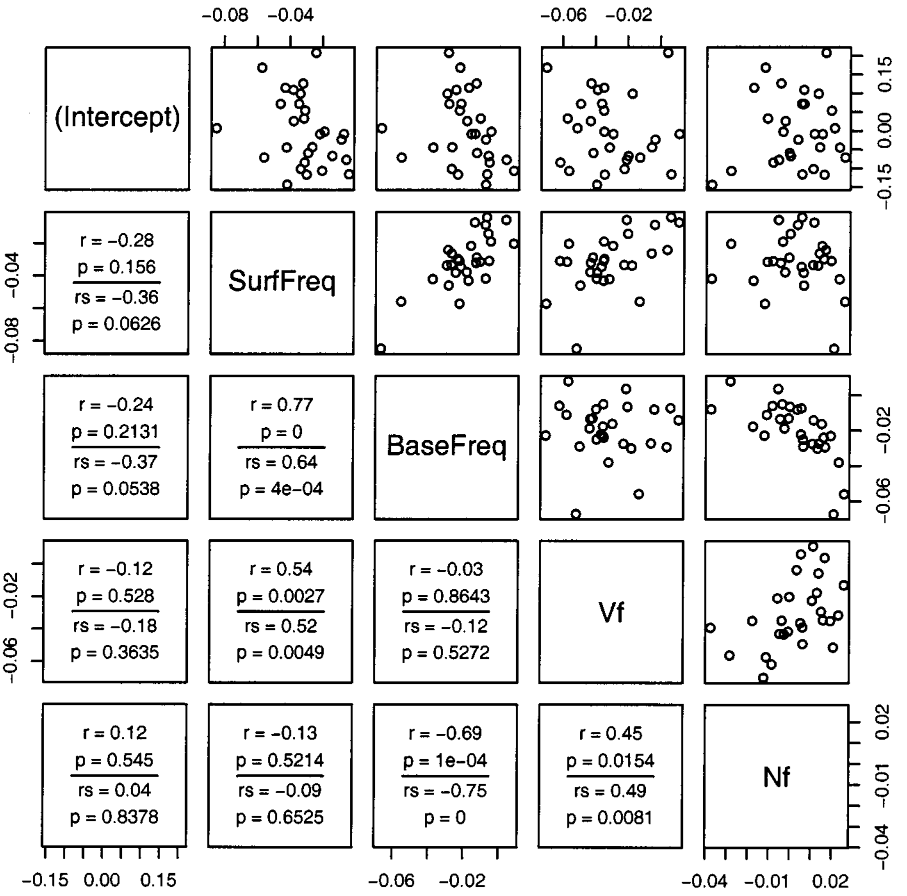

FIG. 4. Pairs plot for the coefficients of separate principal components regression models fit to the responses of individual participants in visual lexical decision. SurfFreq = log Surface Frequency; BaseFreq = log Base Frequency; Vf = log Family Size; Nf = log Family Frequency. With centering; points represent individual participants.

into semantic processing (see, e.g., De Jong et al., 2000), this correlation between the coefficients of Surface Frequency and Family Size suggest that Surface Frequency may have an underlying conceptual component.

A third, also new, correlation emerges for Base Frequency and Family Frequency. Interestingly, this is a negative correlation: Participants for whom Base Frequency has a relatively large facilitatory effect undergo relatively much inhibition from Family Frequency. Apparently, there is a trade-off between sensitivity to Base Frequency and sensitivity to Family Frequency, with facilitation due to sensitivity to the frequency of the word including its inflectional variants going hand in hand with increasing inhibition by the derived words and compounds in which these words occur as constituents. As if the price paid for speeding up the recognition of the base during visual identification is paid for by increasing competition from the access representations of the family members of the base.

Finally, we observe that Family Size and Family Frequency are positively correlated. For large negative coefficients of Family Size, we observe less large negative coefficients for Family Frequency. For small negative coefficients of Family Size, we find positive coefficients for Family Frequency. This suggests that there is a second trade-off between a facilitatory semantic effect of Family Size and an inhibitory competition-driven effect of Family Frequency at the access level.

The corresponding principal components analysis for the centered auditory lexical data, again discarding two principal components with 30.05% of the variation, reveals a significant correlation for the coefficients of Surface Frequency and those of Base Frequency ($r = .81$, $p < .0001$; $r_s = .75$, $p < .0001$), and possibly a negative correlation between Base Frequency and Family Frequency ($r = -.18$, $p = .28$, $r_s = -.33$, $p = .0379$). No other correlations are significant, hence, we do not include a pairs plot for these data.

What these analyses show is that participants may be more of less sensitive to independent variables such as the frequency measures studied here. A greater sensitivity to the facilitatory effect one variable may go hand in hand with greater sensitivity to the facilitatory effect of other variables, as is the case for, e.g., Surface Frequency and Family Size in the visual and auditory modalities. Alternatively, a greater sensitivity to the facilitatory effect of one variable may go hand in hand with a greater inhibitory effect of another variable, as appears to be the case in the visual modality for Family Size and Family Frequency, and Base Frequency and Family Frequency.

GENERALIZING ACROSS PARTICIPANTS

When we fit separate linear models to the responses of individual participants, we are in some way restricting ourselves to the behavior of these particular participants. Even though participants may differ systematically with respect to how sensitive they are to our four frequency measures, we would also like to know to what extent the response latencies of an average subject are predicted by Surface Frequency, Base Frequency, Family Size, and Family Frequency. That is, we need a statistical model that allows us to take into account simultaneously the systematic variation among the participants and the characteristic behavior of a participant with respect to these measures. Linear Mixed Effects (LME) models (Pinheiro & Bates, 2000) meet these requirements.

To understand the LME approach, consider a traditional by-item analysis in which we have, for each of the 36 experimental words, a reaction time (averaged over participants) and four independent variables, Surface Frequency, Base Frequency, Family Size, and Family Frequency. The problem with such a standard by-item analysis is that it tries to come to grips with the overall effect of the independent variables on the processing of the words by averaging over the participants before entering the regression analysis or analysis of variance. This averaging process brings along a tremendous loss of potentially relevant information in the response latencies of individual participants to individual words. By fitting an LME model, we can come to grips with the effects of the independent variables with respect to the average characteristics of the population represented by our participants, and at the same time with the variability among the participants themselves (Pinheiro & Bates, 2000), without first having to invoke averaging over subjects. In matrix notation, the model we will fit to the data is expressed as follows:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \boldsymbol{\varepsilon}, \quad i = 1, \ldots, 28, \\ \mathbf{b}_i &\sim \mathcal{N}_5(\mathbf{0}, \boldsymbol{\Psi}), \ \boldsymbol{\varepsilon} \sim \mathcal{N}_{36}(\mathbf{0}, \sigma^2\mathbf{I}). \end{aligned} \tag{2}$$

We explain the elements of this equation step by step. First, note that i ranges over the 28 subjects of our visual lexical decision experiment. The vector of expected log RTs for subject i is $\mathbf{y}_i$. If subject i has made an error response to word j then a missing value is recorded at y_{ij}. Fortunately, linear mixed effect models with (restricted) maximum likelihood parameter estimation do not depend on the data being fully balanced (see Pinheiro & Bates, 2000, pp. 25–27 for discussion). **X** denotes the data matrix for items $j = 1, \ldots, 36$ as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & \text{SurfFreq}_1 & \text{BaseFreq}_1 & \text{Vf}_1 & \text{Nf}_1 \\ 1 & \text{SurfFreq}_2 & \text{BaseFreq}_2 & \text{Vf}_2 & \text{Nf}_2 \\ 1 & \text{SurfFreq}_3 & \text{BaseFreq}_3 & \text{Vf}_3 & \text{Nf}_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{SurfFreq}_{36} & \text{BaseFreq}_{36} & \text{Vf}_{36} & \text{Nf}_{36} \end{bmatrix} \tag{3}$$

with SurfFreq_j the log Surface Frequency of the jth word, BaseFreq_j the Base Frequency of the jth word, and Vf_j and Nf_j the corresponding log Family Size and Frequency. The vector $\boldsymbol{\beta}$ specifies the coefficients for the Intercept, Surface Frequency, Base Frequency, Family Size, and Family Frequency for what is called the fixed effects part of the model. If we had used a design with four factors, each with two levels (high versus low) for our frequency measures, these factors would have been fixed effects factors, accounted for by means of different weights in $\boldsymbol{\beta}$**.** The continuous regression variables that we consider in this study are likewise accounted for by means of weights in $\boldsymbol{\beta}$**.** Crucially, this vector specifies the average effects of Surface Frequency, Base Frequency, Family Size, and Family Frequency in the population of participants as estimated from the random sample of participants in our experiment.

The fixed effect component ($\mathbf{X}\boldsymbol{\beta}$) of our model provides the ‘‘best guess’’ about the effect of Surface Frequency, Base Frequency, Family Size, and Family Frequency on response latencies when no additional information about the kind of participant is available. For each individual participant in our experiment, however, we can also estimate to what extent the average $\boldsymbol{\beta}$ weights have to be adjusted in order to make the model more precise for these particular participants. This is accomplished by what is called the random effects part of the model ($\mathbf{Z}\mathbf{b}_i$). In the case of our visual data, $\mathbf{Z} = \mathbf{X}$**,** as the analysis using Eq. (1) showed that the participants in the visual lexical decision experiment have differential sensitivity to all four frequency measures. The vector $\mathbf{b}_i$ specifies to what extent the overall fixed effect predictions have to be modified for participant *i.* In other words, $\mathbf{b}_i$ captures the individual sensitivity of a participant to Surface Frequency, Base Frequency, Family Size, and Family Frequency. Another way of putting this is that the $\mathbf{b}_i$ coefficients measure the difference for each participant from the average response to Surface Frequency, Base Frequency, Family Size, and Family Frequency. For the auditory data, **Z** will contain coefficients for the Intercept, Surface Frequency, and Base Frequency only, since the random regression model for the auditory data revealed structure for Surface Frequency and Base Frequency only. The random effects $\mathbf{b}_i$ may have different variances and covariances, as specified by the variance–covariance matrix $\boldsymbol{\Psi}$**,** which in our case is a 5×5 matrix (the intercept and four independent variables). The residual error $\boldsymbol{\varepsilon}$ is identical for all participants and words; **I** is the j by j identity matrix. This completes the definition of our LME model.

From the point of view of a standard by-item analysis, linear mixed effect models can be thought of as an item analysis in which, instead of averaging over subjects, the subjects are brought into the model in a different component or, in statistical terminology, at a different level, the level of the random effects. Note that there are two levels of random variation in this model, the level of the words (items) and the level of the participants (subjects). In this kind of multilevel modeling (see Pinheiro & Bates, 2000, p. 61), separate by-item and by-subject analyses are no longer necessary.

To see the increase in power of linear mixed effect models, we begin with a simple

TABLE 1
Summary of Significant Effects by Type of Analysis

		SurfFreq	BaseFreq	Vf	Nf
Auditory LD	Item analysis (4)				+
	Linear model (5)	+		+	+
	LME (2)	+		+	+
	PC-LME (2)	+	+	+	+
Visual LD	PC-LME (2)	+	+	+	+
Rating	PC-LME (2)	+	+	+	

Note. PC-LME denotes a linear mixed effects principal components regression. Numbers refer to the corresponding equations in the text.

regression analysis based on the item means, obtained by averaging over the responses of the subjects. Table 1 provides a summary overview of the results of this analysis and the analyses that are to follow. We consider the data from the auditory lexical decision experiment. According to the by-item regression analysis, which is based on the model

$$\mathrm{E}[E_i[y_{ij}]] = \beta_0 + \beta_1 \mathrm{Surf\,Freq}_j + \beta_2 \mathrm{Base\,Freq}_j + \beta_3 \mathrm{Vf}_j + \beta_4 \mathrm{Nh}_j, \qquad (4)$$

with $j = 1, 2, \ldots, 36$, there is a significant effect of Family Frequency only [Family Frequency: $F(1, 31) = 7.46$, $p = .0103$; Surface Frequency: $F(1, 31) = 2.0$, $p = .1678$; Base Frequency: $F < 1$; Family Size: $F(1, 31) = 2.05$, $p = .1620$]. Instead of averaging over subjects, we might consider a regression model that uses all 1350 individual responses to the items, but that ignores the by-subject grouping structure:

$$\mathrm{E}[y_k] = \beta_0 + \beta_1 \mathrm{Surf\,Freq}_k + \beta_2 \mathrm{Base\,Freq}_k + \beta_3 \mathrm{Vf}_k + \beta_4 \mathrm{Nf}_k, \qquad (5)$$

with $\mathrm{Surf\,Freq}_k$ denoting the Surface Frequency of the word that appeared in the *k*th word trial in the experiment. This revised regression model reveals considerably more structure. In addition to Family Frequency [$F(1, 1345) = 86.11$, $p < .0001$], Surface Frequency [$F(1, 1345) = 17.95$, $p < .0001$] and Family Size [$F(1, 1345) = 22.37$, $p < .0001$] emerge as significant predictors of RT [Base Frequency: $F(1, 1345) = 1.19$, $p = .2762$]. A much better fit to the experimental data is obtained, however, with a linear mixed effects model as specified above as Eq. (2). As before, Surface Frequency [$F(1, 1307) = 22.96$, $p < .0001$], Family Size [$F(1, 1307) = 31.32$, $p < .0001$], and Family Frequency [$F(1, 1307) = 90.47$, $p < .0001$] are significant predictors [Base Frequency: $F(1, 1307) = 2.09$, $p = .1486$]. In the random effects part of the LME model, the coefficients for the Intercept are very important. By allowing each participant her or his own intercept, i.e., by taking into account that some subjects are, on average, fast, and that others are, on average, slow, the LME model is able to partial out an error component that is confounded in the undifferentiated model [Eq. (5)] with the independent variables. The superiority of the LME model is already visible from an inspection of the mean squared error, .0444 for the undifferentiated regression model [Eq. (5)], and .0341 for the LME model, a reduction of 23%. This informal observation is confirmed formally by the likelihood ratio test (LRT) statistic (Pinheiro & Bates, 2000, p. 83). This statistic is used to compare a more general model with a more restricted model that has the same fixed effects structure. In our case, the LME model, which has 21 parameters, is the more general model. The undifferentiated regression model [Eq. (5)] is the more specific model, as it has only six nonzero parameters. The undifferentiated model [Eq. (5)] is more specific in the sense that it constrains the variances and covariances $\boldsymbol{\Psi}$ in Eq. (2) to be equal to zero. If L_2 is the likelihood of the more general model with k_2 parameters and L_1 the likelihood of the more specific model with k_1 parameters, the LRT statistic:

$$\text{LRT} = 2 \log(L_2/L_1), \tag{6}$$

is asymptotically $\chi^2_{k_2-k_1}$ distributed under the null hypothesis that the restricted model is valid. For the present data, the LRT equals 219.62 ($p < .0001$), confirming that the LME model is superior.

The analyses presented thus far ignore the problem of the collinearity in our data matrix. In order to ascertain that the results we have obtained are not due to collinearity, we conclude our analysis with a linear mixed effects principal components regression. This analysis reveals significant coefficients for all four frequency measures, now including Base Frequency as well: Surface Frequency: $\hat{\beta}_1 = -0.021$, $p < .0001$; Base Frequency: $\hat{\beta}_2 = -0.018$, $p < .0001$; Family Size: $\hat{\beta}_3 = -0.013$, $p < .01$; and Family Frequency: $\hat{\beta}_4 = 0.007$, $p < .001$. For the visual data, we find a very similar pattern of results: Surface Frequency: $\hat{\beta}_1 = -0.031$, $p < .0001$; Base Frequency: $\hat{\beta}_2 = -0.020$, $p < .0001$; Family Size: $\hat{\beta}_3 = -0.033$, $p < .0001$; Family Frequency: $\hat{\beta}_4 = 0.004$, $p < .04$. The emergence of Base Frequency as a significant predictor is in line with the random regression model according to which participants are differentially sensitive to Base Frequency.

In these regression analyses, Family Size has a negative coefficient, indicating facilitation, while Family Frequency has a positive coefficient, indicating inhibition. We interpret the inhibitory effect of Family Frequency as an effect arising at the level of form identification due to lexical competition between access representations. We interpret the facilitatory effect of Family Size as a semantic effect arising due to semantic activation spreading between morphologically related words (De Jong et al., 2000). Interestingly, the coefficient for Family Frequency in the auditory modality is larger than the corresponding coefficient in the visual modality, while at the same time the coefficient for Family Size less negative, suggesting that perhaps inhibitory lexical competition effects (Meunier & Segui, 1999) at the access level are more severe in the auditory modality than in the visual modality.

Finally, we carried out a linear mixed effects principal components regression analysis on the subjective frequency rating data that we have available for our set of words (cf. Schreuder & Baayen, 1997). Surface Frequency ($\hat{\beta}_1 = .349$, $p < .0001$), Base Frequency ($\hat{\beta}_2 = .164$, $p < .0001$), and Family Size ($\hat{\beta}_3 = .494$, $p < .0001$), but not Family Frequency ($\hat{\beta}_4 = .006$, $p = .662$), emerge from this analysis as significant predictors. The coefficients are positive: A higher Surface Frequency leads to a shorter reaction time but to a higher subjective frequency rating. The absence of an effect of Family Frequency for a subjective frequency rating task makes sense, as there is no time pressure enforcing rapid identification. By the time participants write down their rating, effects arising during the early stages of form identification are no longer visible.

CONCLUSIONS

This methodological study has shown that traditional by-subject and by-item analyses ignore systematic variation in the way individual participants respond to independent variables such as Surface Frequency, Base Frequency, Family Size, and Family Frequency in our data. This study has also shown that the by-subject and by-item analyses are not powerful enough to reveal all the significant effects present in our data, due to information loss by prior averaging. By using linear mixed effect models on the complete data matrix, we have been able to resolve inconsistencies between results obtained for the visual and auditory modalities and to obtain a more profound insight in the role of competition of morphologically related words at the form level and the role of facilitation at the level of semantic processing.

Further research will have to clarify how the variation in the behavior of a single participant at different moments in time relates to the interparticipant variation documented in the present study. Our intuition, however, is that the intraparticipant differences will turn out to be less than the interparticipant variation.

To our knowledge, psycholinguistics is the only domain of scientific inquiry in which by-subject and by-item analyses form the gold standard of experimental statistics. We think that the present results show that it is worthwhile to consider using more powerful techniques that have been developed in the past decade.

REFERENCES

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity* (Wiley Series in Probability and Mathematical Statistics). New York: Wiley.

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language,* **42,** 390–405.

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: Memory, Learning, and Cognition,* **26,** 419–511.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes,* **15,** 329–365.

Hocking, R. R. (1996). *Methods and applications of linear models. Regression and the analysis of variance.* New York: Wiley.

Meunier, F., & Segui, J. (1999). Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language,* **41,** 327–344.

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

Rubenstein, H., & Pollack, I. (1963). Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior,* **2,** 147–158.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance,* **3,** 1–17.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language,* **37,** 118–139.

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