Behavior Research Methods
2007, 39 (1), 101-117

Multilevel models for the experimental psychologist: Foundations and illustrative examples

Lesa Hoffman and Michael J. Rovine
Pennsylvania State University, State College, Pennsylvania

Although common in the educational and developmental areas, multilevel models are not often utilized in the analysis of data from experimental designs. This article illustrates how multilevel models can be useful with two examples from experimental designs with repeated measurements not involving time. One example demonstrates how to properly examine independent variables for experimental stimuli or individuals that are categorical, continuous, or semicontinuous in the presence of missing data. The second example demonstrates how response times and error rates can be modeled simultaneously within a multivariate model in order to examine speed–accuracy trade-offs at the experimental-condition and individual levels, as well as to examine differences in the magnitude of effects across outcomes. SPSS and SAS syntax for the examples are available electronically.

One of the most important pieces in the toolbox of the experimental psychologist is the ANOVA model. ANOVA models are well suited to an analysis of the impact on a continuous response variable of categorical design factors (independent variables) that are manipulated or measured between subjects, within subjects, or some combination of both (e.g., split-plot). Examples of such categorical design factors include the number of items held in memory during completion of a second task (e.g., 3, 6, or 9 items) and the types of distractors surrounding a visual target (e.g., none, similar, dissimilar). For many investigators, ANOVA models are more than adequate to examine the research hypotheses of interest from their experimental design. However, in other instances, ANOVA models may not be appropriate. For example, although ANOVA models can be extended in order to examine the main effect of continuous person-level covariates such as age or ability, the analysis of covariance (ANCOVA) model is only appropriate if interactions between the categorical design factors and continuous covariates do not exist (i.e., the assumption of homogeneity of regression). In some applications, however, such interactions may very well be the focus of interest (e.g., the extent to which the effects of memory load or type of distractor vary across age or ability levels).

The matter may be further complicated in the case of continuous within-subjects factors. In real-world experimental stimuli such as photographs, text passages, or autobiographical memories, the design features of interest (e.g., visual complexity of the photograph, difficulty of the text passage, or strength of the memory) must be measured instead of manipulated. As a result, these experimental stimuli may vary continuously in their levels of a design feature, just as persons may vary continuously in characteristics or abilities. Manipulated variables (e.g., dosage levels) may also be continuous. What if an interaction between a continuous person variable and a continuous design factor were of substantive interest? Such interactions of continuous between-subjects design factors or person variables can readily be examined within a general linear modeling framework using multiple regression, of which between-groups ANOVA is a special example.

If the design factor were administered within subjects instead, however, there would be fewer options for examining its main effect and its interaction with continuous person-level covariates. An all too common solution to this dilemma is to categorize the continuous independent variables (either stimulus-level design factors or subject-level individual-difference variables) in order to fit them within an ANOVA model. However, because the categorization of continuous independent variables substantially reduces the power to detect effects and inflates Type I error rates, methodologists strongly discourage doing so (e.g., Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993).

Alternative approaches for analyzing repeated measures data with continuous design factors have made use of variations on linear regression methods. Although typical regression models cannot be used on the pooled data set of within-subjects data due to violation of the assumption of independence (i.e., model residuals from the same person may be more related than those from different

L. Hoffman, lhoffman2@unlnotes.unl.edu

 Copyright 2007 Psychonomic Society, Inc.

people), several methods for circumventing this problem have been suggested. One alternative is known as *fixed effects regression* (see Allison, 1994; Lorch & Myers, 1990; Snijders & Bosker, 1999, pp. 41–45), in which $n-1$ dummy indicator variables for n persons and $n-1$ person × design feature interaction variables are included in order to control for any within-subjects residual correlation. Because the dummy indicator variables will account for all of the between-subjects differences, however, a significant limitation of this approach is that no other person-level independent variables can be examined within the model. Additionally, this approach draws no inferences from a population of individuals, which is often in contrast with the intentions of the analyst, who may indeed wish to generalize to other samples.

A second alternative is a two-stage approach known as *slopes as outcomes* (see Lorch & Myers, 1990; Singer & Willett, 2003, pp. 28–44), in which regressions are performed separately for each person in the first step, and the individual regression estimates are then used as data in a between-subjects analysis (i.e., ANOVA or regression). Although intuitively appealing, this method does not account for the differential reliability of the individual regression estimates, which can result in biases in unknown directions. Such two-stage procedures are also statistically inefficient and are generally not recommended (Singer & Willett, 2003; Snijders & Bosker, 1999).

A third alternative is the univariate approach to repeated measures using modified error terms within a general linear model framework, in which the significance of effects is assessed using customized error terms that properly account for between-subjects variation (see Lorch & Myers, 1990; O'Brien & Kaiser, 1985; Rovine & von Eye, 1991, pp. 26–28). The selection of the correct error term for a given contrast can be challenging for a less sophisticated user, and there are two significant limitations to the univariate approach given that it is based on least squares estimation: (1) It assumes a particular pattern of variances and covariances, and (2) it assumes that data are missing completely at random. These limitations will be discussed later in greater detail.

Although not commonly used in experimental psychology, state of the art multilevel modeling approaches often used in other disciplines represent a viable alternative to ANOVA or regression-based approaches for repeated measures designs. The purpose of this article is to illustrate how multilevel models can fit into the toolbox of the experimental psychologist in order to answer substantive questions about design features that simply don't fit within traditional repeated measures models. Multilevel models (MLMs, also known as hierarchical linear, random coefficients, or general linear mixed models; Laird & Ware, 1982) are often used in the literature of educational, family, developmental, and organizational psychology to analyze data in which there are sources of nesting, and for which assumptions of independence are likely to be violated. For example, students from the same school, members of the same family, and people in the same organization may be more alike in their responses than people from different schools, families, or organizations. In the developmental literature, multilevel models are often used to examine individual differences in change over time, where time points are nested within individuals (i.e., growth curve models). These higher order groupings are specified as varying randomly from one another, however, not treated as fixed; thus, predictors of this random variation between higher order units, as well as within higher order units, may be evaluated explicitly.

What may not be immediately obvious is how experimental stimuli such as trials or items can also be nested within individuals (i.e., in designs in which only certain individuals receive certain items), or crossed with individuals (i.e., in designs in which every individual receives every item). In this article, the foundations of the multilevel model as it relates to more familiar ANOVA and regression models will be presented as it applies to analysis of data from experimental designs, along with two illustrative examples. For a technically rigorous treatment, the reader is invited to consult one of the many excellent texts dealing with multilevel models in the clustered or nested cases (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and in the growth-curve cases (Fitzmaurice, Laird, & Ware, 2004; Singer & Willett, 2003). Although many excellent MLM tutorials are also available (Diez-Roux, 2000; Quené & van den Bergh, 2004; Sayer & Klute, 2004; Singer, 1998), the present article differs from them in two respects: (1) Our focus is on the specific advantages of the multilevel model for use with experimental designs, as discussed in greater detail below; and (2) our exposition is designed to be accessible to researchers familiar only with ANOVA and regression. As a result, we think the detailed presentation of these methods within a familiar context, as well as the availability of example syntax and data in electronic appendices (see the Author Note at the end of this article), will help to facilitate adoption of these methods by interested experimental psychologists. Estimation of multilevel models is now widely available within popular software packages such as SPSS, SAS, HLM, MLwiN, and Mplus. Some, such as SAS and Mplus, are syntax-based, and some—SPSS, HLM, MLwiN—are Windows based (although syntax may also be used in some of the latter packages). These packages also differ in how the model is programmed, with SPSS and SAS implementing the general linear mixed model as a single equation and the others doing so as multilevel equations. The more intuitive multilevel equation presentation is used here.

ADVANTAGES OF THE MULTILEVEL MODEL FOR EXPERIMENTAL DESIGNS

The multilevel model can be conceptualized as a series of interrelated regression models that explain sources of variance at multiple levels of analysis, such as at the experimental stimuli and person levels. As will be explained in further detail, one of the hallmarks of the multilevel model is its distinction between *fixed effects* and *random effects*. Fixed effects are most familiar to general users, and are effects of variables that are specified as constant, or *fixed*, over all individuals in the sample (e.g., regression weights, mean differences). In contrast, random effects are

effects of variables that are specified as *varying* over all individuals in the sample. As will be shown, the repeated measures ANOVA model is merely a restricted version of the multilevel or general linear mixed model. The removal of these restrictions has the following advantages for the analysis of data from experimental designs:

1. Great flexibility is possible in addressing dependencies among observations (i.e., correlated residuals) with alternative covariance structures or random effects.
2. Main effects and interactions of categorical, continuous, or semicontinuous independent variables for stimuli or for individuals may be examined simultaneously.
3. Listwise deletion is not required; data from individuals with only partial response (by accident or by design) can still be included in the model to maximize power.
4. Multivariate models can be used in order to achieve greater power in testing fixed effects, to examine differences in fixed effects across response variables, and to examine correlations among response variables at the stimuli or individual levels.

Let us consider as background for our discussion an example experiment in which 50 observers (denoted by i) are each presented with 30 sentences (denoted by t), and the speed with which the sentences are read aloud is the outcome measure. The predictors that pertain to the sentences are active versus passive voice (scores of 0 or 1; denoted as V_{ti}) and syntactic complexity (continuous scores of 1 to 20; denoted as C_{ti}). The predictor that pertains to the individuals is verbal fluency (continuous scores of 10 to 50; denoted as F_i). In order to fit these data into an ANOVA model, one might collapse sentence complexity and verbal fluency each into categories of low (0) or high (1). (Note that this is done here for pedagogical purposes, and is not recommended.) The split-plot ANOVA model in multilevel form is shown in Equation 1:

$$\begin{aligned} \text{Level 1:} \quad y_{ti} &= \beta_{0i} + \beta_{1i}(V_{ti}) + \beta_{2i}(C_{ti}) \\ &\quad + \beta_{3i}(V_{ti})(C_{ti}) + e_{ti} \\ \text{Level 2:} \quad \beta_{0i} &= \gamma_{00} + \gamma_{01}(F_i) + U_{0i}, \\ \beta_{1i} &= \gamma_{10},\ \beta_{2i} = \gamma_{20},\ \beta_{3i} = \gamma_{30}, \end{aligned} \tag{1}$$

where y_{ti} is the observed reading time and e_{ti} is the residual (i.e., the difference between observed and model-predicted reading time) for sentence t and individual i. The Level 1 residuals (e_{ti}) are assumed to be normally distributed overall and with constant variance across the sentences. All 1,500 potential reading times (i.e., 30 sentences × 50 individuals) are modeled simultaneously. The Level 1 model describes the relation between each reading time and the sentence predictors. The effects of the sentence predictors (the β_is) are then themselves outcomes for each subject in each equation of the Level 2 model.

Fixed effects are denoted with γs: γ_{00} is the fixed intercept, or the expected reading time for a sentence of active voice and low complexity for a person of low fluency (i.e., when V_{ti}, C_{ti}, and $F_i = 0$), and γ_{10} and γ_{20} are the fixed (main) effects of the sentence predictors, or the mean difference of active versus passive voice (when $C_{ti} = 0$) and low versus high complexity (when $V_{ti} = 0$), and γ_{30} is the fixed effect for the voice by complexity interaction, or the expected additional effect on reading time when voice is passive and complexity is high (i.e., when $V_{ti} = 1$ and $C_{ti} = 1$). Note that the effects of voice and complexity in the Level 2 model (γ_{10}, γ_{20}, and γ_{30}) are replaced directly by β_{1i}, β_{2i}, and β_{3i} in the Level 1 model. This implies that the main effects of voice and complexity and their two-way interaction are expected to be the same across individuals, the definition of a fixed effect. In contrast, the Level 2 model for the intercept (β_{0i}) contains two terms besides the fixed intercept (γ_{01}): γ_{01}, the fixed (main) effect for fluency, or the mean difference between low and high fluency (i.e., when $F_i = 1$), and U_{0i}, the individual random intercept, or individual-specific deviation from the fixed intercept.

It is important to discuss at this point the implications of including all observations (i.e., 30 sentences × 50 individuals) within the same model. In a typical ANOVA, observations within the same condition are averaged and these condition means then analyzed. This procedure implicitly considers the sentences to be fixed effects; that is, variation in reading time due to systematic differences among sentences within the same condition is removed prior to analysis (see Raaijmakers, Schrijnemakers, & Gremmen, 1999, for an extended discussion). Rather than artificially removing that sentence variability, however, in this example it is retained in the analysis but must be incorporated specifically into the model. One way in which to address the systematic effect of sentence on reading time that remains after accounting for the effects of voice and complexity is to include a random effect for sentence, as in Equation 2:

$$\begin{aligned} \text{Level 1:} \quad y_{ti} &= \beta_{0i} + \beta_{1i}(V_{ti}) + \beta_{2i}(C_{ti}) \\ &\quad + \beta_{3i}(V_{ti})(C_{ti}) + W_t + e_{ti} \\ \text{Level 2:} \quad \beta_{0i} &= \gamma_{00} + \gamma_{01}(F_i) + U_{0i}, \\ \beta_{1i} &= \gamma_{10},\ \beta_{2i} = \gamma_{20},\ \beta_{3i} = \gamma_{30}, \end{aligned} \tag{2}$$

where all parameters are as in Equation 1, and the new parameter W_t is the random effect for sentence. Because each individual was presented with each sentence, sentences are actually crossed with individuals at Level 2, such that each trial (i.e., sentence × subject combination) is nested within sentences and within subjects. If each individual had received a different sentence (e.g., if individuals each had written their own sentences), then sentences would be strictly nested within individuals, rather than crossed with individuals at Level 2, as in this example. For convenience the random sentence effect is included directly in the Level 1 model, rather than in its own Level 2 equation. Each reading time is thus modeled as a function of the fixed effects of sentence type (voice, complexity, and their interaction), the fixed effect of fluency, the random effect of individual i, and the random effect of sentence t. The trial-to-trial variation that remains after accounting for the systematic effects of sentences and of individuals is represented by e_{ti}.

The advantages of the multilevel model for the analysis of experimental designs as outlined above will now be

presented in greater detail as they relate to the previous example.

Dependencies Among Observations

Alternative covariance structures. In a typical ANOVA, items are averaged into condition means (e.g., for voice by complexity), which are then subjected to analysis. One of the assumptions of this ANOVA model is that individuals differ in only one way (e.g., in overall reading times). This implies that the residual variance within condition (as well as the covariances between the residuals from each condition) should be equal after controlling for the random intercepts, a condition known as compound symmetry, as shown in the first part of Table 1. Compound symmetry is slightly more restrictive than the condition of sphericity, in which the variances and covariances of orthogonal contrasts of the original repeated measures are assumed to be equal (Huynh & Feldt, 1980). When sphericity does not hold (i.e., when residual variances are larger in some conditions than in others, or more related across some conditions than others), then tests of the fixed effects from the ANOVA model may be incorrect.

An alternative is the multivariate approach to repeated measures ANOVA, in which the orthogonal contrasts are analyzed simultaneously, and in which no assumptions are made regarding the structure of the residual variance–covariance matrix (analogous to all variances and covariances being estimated separately; i.e., an *unstructured* matrix, as seen in the second part of Table 1). Thus, rather than assuming a common error term for all fixed effect comparisons, a condition-specific error term is used for each separate contrast. This results in greater power for each univariate test, but can result in less power for the overall multivariate test when compared to an omnibus test adjusted for the degree of violation of sphericity (Maxwell & Delaney, 2003).

Multilevel models—or general linear mixed models, as they are often referred to in this context—can be used as alternatives to ANOVA when the assumption of sphericity is likely to be violated (e.g., Littell, Pendergast, & Natarajan, 2000; Maas & Snijders, 2003; Wallace & Green, 2002), because they have been shown to have greater power in detecting fixed effects than ANOVA models when conditions of sphericity are not met (Quené & van den Bergh, 2004). Multilevel models can also provide a useful compromise between the nonparsimonious option of estimating all possible residual variances and covariances—the multivariate approach—and the overly-restrictive option of assuming sphericity—the univariate approach. One such alternative is compound symmetry with heterogeneous variances, as seen in the third part of Table 1, which allows unequal residual variances across conditions but still assumes the correlation among the residuals to be the same across conditions. An advantage of multilevel models over ANOVA

Table 1
Alternative Structures of the Residual Variances and Covariances in Multilevel Models

Type of Model	Variance–Covariance Structure Across Low/High Voice by Low/High Complexity Conditions
Univariate repeated measures ANOVA: *Compound symmetry*	$\begin{bmatrix} v & & & \\ r & v & & \\ r & r & v & \\ r & r & r & v \end{bmatrix}$
Multivariate repeated measures ANOVA and mixed linear model: *Unstructured*	$\begin{bmatrix} v_{11} & & & \\ r_{21} & v_{22} & & \\ r_{31} & r_{32} & v_{33} & \\ r_{41} & r_{42} & r_{43} & v_{44} \end{bmatrix}$
Example alternative structure from the mixed linear model: *Compound symmetric, heterogeneous variances*	$\begin{bmatrix} v_1 & & & \\ r & v_2 & & \\ r & r & v_3 & \\ r & r & r & v_4 \end{bmatrix}$
Mixed linear model: *Random effects* (*G; between-subjects unstructured form) and residual* (*within-subjects R; identity form) variance matrices*	G Matrix: $\begin{pmatrix} v_{U_0} & & \\ r_{21} & v_{U_1} & \\ r_{31} & r_{32} & v_{U_2} \end{pmatrix}$ R Matrix: $\left(\begin{array}{cccc\|cccc} v_{res} & & & & & & & \\ 0 & v_{res} & & & & & & \\ 0 & 0 & v_{res} & & & & & \\ 0 & 0 & 0 & v_{res} & & & & \\ \hline 0 & 0 & 0 & 0 & v_{res} & & & \\ 0 & 0 & 0 & 0 & 0 & v_{res} & & \\ 0 & 0 & 0 & 0 & 0 & 0 & v_{res} & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_{res} \end{array}\right)$

models is that one need not make any assumptions about the structure of the residual variances and covariances. A variety of alternative structures can be estimated and their fit compared empirically in order to ensure the most appropriate tests of the fixed effects. It is also possible to estimate separate residual variance–covariance matrices for different values of a person-level predictor (see Littell, Milliken, Stroup, & Wolfinger, 1996).

Random effects. The direct specification of an alternative structure for the residual variance–covariance matrix is one way to account for variances and covariances that differ across conditions. Yet when the source of the heterogeneity across conditions is thought to arise from individual differences in a meaningful process, another variant of the multilevel model may be more useful instead in accounting for the dependency among observations: the random effects model, as seen in the bottom part of Table 1. This model can be estimated without requiring any averaging into condition means. In a random effects model, heterogeneity of the variances and covariances is modeled by two matrices: one matrix of random effects (the G matrix; here, a random intercept and random effects for sentence voice and complexity, as described below), and one matrix for the residuals (the R matrix), which are assumed to have constant variance and be uncorrelated across individuals and observations after accounting for the random effects. As with alternative structures for the residual variance–covariance matrix, separate random effects matrices can also be estimated for different values of person-level predictors, as warranted.

The ANOVA model given in Equation 1 is also known as a random intercept model, given that the individual intercepts (β_{0i}) were comprised of the sample intercept (fixed effect γ_{00}) and the person-specific random deviations (U_{0i}) from the fixed intercept. Because the effects of sentence voice and complexity were assumed to be fixed, any differences among subjects in the magnitude of these effects are considered residual error. Thus, to the extent that individuals differ systematically in the extent to which their reading times vary by sentence voice or complexity, the ANOVA model will not be appropriate. Such a restriction is not required in the multilevel model, of which the repeated measures ANOVA model is merely a special case. The restriction of fixed effects only for sentence voice and complexity is relaxed in Equation 3:

$$\begin{aligned}
\text{Level 1:} \quad y_{ti} &= \beta_{0i} + \beta_{1i}\left(V_{ti}\right) + \beta_{2i}\left(C_{ti}\right) \\
&\quad + \beta_{3i}\left(V_{ti}\right)\left(C_{ti}\right) + W_t + e_{ti} \\
\text{Level 2:} \quad \beta_{0i} &= \gamma_{00} + \gamma_{01}\left(F_i\right) + U_{0i}, \\
\beta_{1i} &= \gamma_{10} + U_{1i},\ \beta_{2i} = \gamma_{20} + U_{2i}, \\
\beta_{3i} &= \gamma_{30}, \qquad (3)
\end{aligned}$$

where all terms are as in Equation 2, except that the individual effects of sentence voice (β_{1i}) and complexity (β_{2i}) now comprise the fixed effects (γ_{10} and γ_{20}) as well as person-specific random effects (U_{1i} and U_{2i}), or deviations from the fixed effects. In other words, subjects are permitted to vary systematically from one another in the magnitude of their response to sentence voice and complexity. By convention, random effects are not estimated for the interaction of voice × complexity, but instead are estimated only for their main effects. The random effects are assumed to have a multivariate normal distribution across individuals. It is important to note that random variation over higher level units (i.e., if individuals are themselves nested in groups) can also be accommodated as a multilevel model with three or more levels.

In repeated measures ANOVA, the random intercepts are modeled directly as differences across persons in their overall level. Their variance is then partialed out of the error terms used in the *F* tests, but is otherwise not of direct interest. In contrast, in the multilevel model, rather than estimating the random effects directly (for the individual intercepts, as well as for effects of other predictors or for the sentences), the magnitude of the variance of the random effects is estimated instead, and the random effects can then be predicted after the fact, on the basis of the model.

Two questions are relevant for each individual random effect: (1) Is the variance of the random effect significant? That is, does the size of the effect differ systematically among individuals, or should it instead be considered fixed across individuals? and (2) To what extent can the variance of the individual random effects be reduced by including individual-level predictors in the model? The parameters for the individual random effects are themselves outcomes (i.e., are error variances) at Level 2. Similarly, the parameters for the random sentence effects are also outcomes at Level 2. That is, just as there is a single error variance to be reduced by predictor variables within regression, similarly, there are multiple such error variances (i.e., individual random effects and random sentence effects at Level 2, trial-to-trial residual variance at Level 1) to be reduced by predictors at each level in a multilevel model. This partitioning of the total variance in the outcome (e.g., reading times) has direct implications for the kinds of predictor variables that can be examined within the model, as described next.

Multilevel Model Specification of Fixed Effects

Unlike the general linear model in which there is a single error term to be reduced, the multilevel model can make it easier to examine the effects of predictors at multiple levels of analysis, because separate error variances are specified at each level. Thus, the inclusion of sentence-level predictors (e.g., voice and complexity) serves to reduce the random sentence variance, and the inclusion of individual-level predictors (e.g., verbal fluency) serves to reduce the individual random-effects variance. However, the multilevel model is similar to the general linear model, in that it allows tests of both main effects and interactions among predictors that are categorical, continuous, or semicontinuous (i.e., piecewise linear effects). The result of such flexibility is that the distorted, dichotomous versions of sentence complexity and verbal fluency that have been used thus far are no longer necessary. Instead of dummy variables for *low* or *high*, the predictors are included in the model in their original continuous metric,

but were centered by subtracting a constant of 10 from complexity (with a range of 1 to 20) and a constant of 30 from fluency (with a range of 10 to 50) for reasons explained below. The model in Equation 3 can be modified to include continuous predictors and their interactions, as shown in Equation 4:

$$\begin{aligned}
\text{Level 1:}\quad y_{ti} &= \beta_{0i} + \beta_{1i}\left(V_{ti}\right) + \beta_{2i}\left(C_{ti} - 10\right) \\
&\quad + \beta_{3i}\left(V_{ti}\right)\left(C_{ti} - 10\right) + W_t + e_{ti} \\
\text{Level 2:}\quad \beta_{0i} &= \gamma_{00} + \gamma_{01}\left(F_i - 30\right) + U_{0i}, \\
\beta_{1i} &= \gamma_{10} + \gamma_{11}\left(F_i - 30\right) + U_{1i}, \\
\beta_{2i} &= \gamma_{20} + \gamma_{21}\left(F_i - 30\right) + U_{2i}, \\
\beta_{3i} &= \gamma_{30} + \gamma_{31}\left(F_i - 30\right), \qquad (4)
\end{aligned}$$

where y_{ti} and e_{ti} still represent the observed reading time and residual error for individual i and sentence t. However, the individual intercept (β_{0i}) now represents the expected reading time for a sentence of active voice and moderate complexity (i.e., $V_{ti} = 0$ and $C_{ti} - 10 = 0$) for a person of moderate verbal fluency (i.e., $F_i - 30 = 0$). It is important to note that the location of the intercept is arbitrary within any model, and its interpretation can often be facilitated by centering any continuous predictors, as we have done here, by subtracting a constant in order to place the origin within the observed range of the variable. For example, if the variable for fluency with an observed range of 10 to 50 were included as is, the intercept would represent the expected reading time for someone with a fluency score of 0, which is not possible given the scale of the variable. By subtracting a constant (e.g., the sample mean) from each individual's fluency score, the scale of the predictors is shifted, such that the intercept then represents the expected reading time for an individual of average fluency. Any constant within the range of the predictor could be used as a centering point, but the mean is commonly used for ease of interpretation. See Kreft, de Leeuw, and Aiken (1995), or Snijders and Bosker (1999) for a more thorough discussion of centering.

In Equation 4, the fixed (main) effect for sentence voice (γ_{10}) refers to the mean difference between active and passive voice (when both $C_{ti} = 0$ and $F_i = 0$). However, the fixed (main) effect for sentence complexity (γ_{20}) now represents a one-unit change in expected reading time for a one-unit change in complexity (when both $V_{ti} = 0$ and $F_i = 0$); that is, γ_{20} is a regression slope. The fixed effect for the voice × complexity interaction (γ_{30}) now represents the expected difference in the size of the complexity slope (when $F_i = 0$) when reading sentences written in the passive voice instead of the active voice—or, similarly, the expected change in the difference between active and passive voice for a one-unit change in complexity (also when $F_i = 0$). The fixed (main) effect of verbal fluency (γ_{01}) now represents a one-unit change in the intercept for a one-unit change in fluency. The fixed effects for the interactions of voice × fluency (γ_{11}), complexity × fluency (γ_{21}), and voice × complexity × fluency (γ_{31}) represent one-unit changes in the effects of voice, complexity, and voice × complexity for a one-unit change in fluency. Thus, the main effects of sentence voice and complexity (β_{1i} and β_{2i}) are now a function of the overall fixed effects (γ_{10} and γ_{20}), the effects of verbal fluency (γ_{11} and γ_{21}), and individual-specific random effects (U_{1i} and U_{2i}). In other words, although individuals are allowed to vary randomly in their overall level for reading time and in the extent to which their reading times are systematically affected by sentence voice and complexity, these random effects for sentence voice and complexity are predicted in part by individual differences in verbal fluency. Further, although the voice × complexity interaction is not considered random, the effect of fluency on the two-way interaction can still be evaluated. Finally, sentences are allowed to vary randomly (W_t) after accounting for the effects of voice and complexity.

Incomplete Responses

Thus far, we have assumed that all possible reading times, 30 sentences × 50 individuals, are included in the model. However, this need not be the case. Incomplete data, one of the greatest challenges to any researcher, can arise in longitudinal studies because of attrition or variable measurement occasions, and within experimental studies can also result from observer fatigue or equipment failure. In these cases, because a repeated measures ANOVA requires complete data, individuals providing partial responses across stimuli cannot be included. Such listwise deletion has long been known to result in reduced power to detect effects (i.e., a loss of efficiency), as well as potential bias in the estimates if the incomplete responses are not missing completely at random (Schafer, 1997). The latter scenario may be particularly likely in certain experimental studies, as when the accuracy of response time data is below ceiling. If incorrect responses are more likely for more difficult items, and response times for incorrect responses are not included (as they almost never are), the response time distribution may no longer be representative, because the highest response times—those for the more difficult stimuli—are likely to be missing. Collapsing across stimuli into condition means (in which different numbers of stimuli are included for each individual) serves only to mask the problem.

The multilevel model addresses missing data by using full-information maximum likelihood to estimate model parameters reflective of those parameters that would have been observed if the data were complete. Maximum likelihood estimation has been shown to provide unbiased and efficient estimates when the data are missing at random, or when the probability of missingness is not related to what the outcome would have been, once predictors related to the missingness are in the model. Thus, rather than eliminating incomplete cases or assuming that missing responses are representative of the distribution of responses, as is required in ANOVA, one can estimate a multilevel model using all available data. Although the assumption of *missing at random* cannot be formally tested, the inclusion of all stimulus- or individual-level predictors (as well as other responses from the individual) should help to obtain the most accurate estimates possible. The assumption of missing at random is also likely to be satisfied when data are incomplete by design, a situation called *planned missingness*, in which different combinations of stimuli are randomly assigned to all individuals.

Schafer (1997) and Schafer and Graham (2002) provide a more thorough treatment of issues in incomplete data.

Just as multivariate versions of the general linear model can be used to analyze multiple outcomes simultaneously, so can a multilevel model, as described below.

Multivariate Models

The multilevel models discussed thus far have been univariate models, in that only one outcome variable (e.g., reading times) has been modeled at a time; however, the multilevel model can be extended to the multivariate case, so that the effects of stimuli-level or individual-level predictors can be tested on multiple outcome variables simultaneously. Multivariate multilevel models have the following advantages over univariate multilevel models (Snijders & Bosker, 1999): First, if the outcomes are correlated, tests of the fixed effects of predictors on each outcome will be more powerful in a multivariate model than the same tests in a univariate model, particularly if the outcomes have incomplete data.

Second, the multivariate test of the effect of a predictor on all outcomes (which can help to reduce Type I error compared to performing separate tests for each outcome) is only possible within a multivariate model. Note that a multivariate test of the predictors requires that the outcomes be on a common scale, since the coefficients are in an unstandardized metric. Transformation of the metric of the dependent variables (e.g., to *z*-scores) may be required in order to perform multivariate tests of fixed effects, although the metrics need not be the same if multivariate tests are not of interest.

Third, one can test hypotheses regarding the differences in magnitude of the effects of the predictors across outcomes. For example, let us assume that our experiment also monitored sentence reading with an eyetracker, so that reading time and total number of fixations for each sentence were both outcome variables of interest. One might conduct two sets of analyses, one for reading times and one for number of fixations, in order to examine the effects on each outcome of sentence voice, sentence complexity, and individual verbal fluency. Although they would reveal whether or not each effect was significant for each outcome, these separate analyses would not reveal whether the predictors had a larger effect on reading times than on number of fixations, or vice versa. For example, if the effect of sentence complexity is significant for reading times but not for number of fixations, whether the magnitude of the complexity effect (i.e., the effect size for complexity) is significantly different across outcomes is optimally tested within a multivariate model. Such comparisons of effect sizes across outcomes are often of interest in experimental studies.

Finally, the multivariate model can be used to examine correlations across outcomes at multiple levels of analysis. Specifically, at the between-subjects level of analysis (Level 2), individual random effects for the intercept and other predictors can be estimated for each outcome, and their covariance can be estimated directly within the multivariate model. This can be useful in examining how much someone who shows a greater than average effect of a given predictor on one outcome is more likely to show a greater than average effect of that predictor on another outcome, as well. At the between-item level of analysis (crossed at Level 2), random item effects for each outcome and their covariances can be estimated in order to examine the extent to which the item deviations are related across outcomes. Finally, at the within-subjects trial level of analysis (Level 1), the estimated covariance among the residuals for each outcome reflects the extent to which response patterns are similar across trials, after controlling for the systematic effects of the predictors, the persons, and the items. In designs without crossed random effects, the multivariate analysis simplifies to between- and within-subjects levels only.

Two in-depth examples are presented in the following section. In the first example, univariate multilevel models for items crossed with individuals are estimated in order to illustrate how to examine the effects of continuous and semicontinuous predictors at multiple levels of analysis, as well as how to accommodate differences in the magnitude of variation across groups. In the second example, multivariate multilevel models (i.e., for experimental conditions nested within individuals) are estimated in order to examine differences in the magnitude of the effects of predictors on response times versus error rates, as well as to examine the possibility of speed–accuracy trade-offs at multiple levels of analysis.

TWO ILLUSTRATIVE EXAMPLES

Example 1: Continuous and Semicontinuous Effects of Items and Persons

Research design. Example 1 was taken from a study that examined the speed with which changes to digital photographs of driving scenes were detected by younger and older adults (Hoffman & Atchley, 2001). Scenes (items) were presented within the flicker paradigm (Rensink, O'Regan, & Clark, 1997), in which original (A) and modified (A′) digital photographs are presented for 280 msec, and blank screens are interspersed for 80 msec. In this presentation (*A–blank–A–blank–A′–blank–A′–blank*, etc.), search for a change between repeated presentations of an otherwise identical scene must be conducted through controlled attentional processing, because local luminance cues at the change location are unable to direct attention in the presence of a global luminance change (the blank screen). Each item was presented for 60 sec or until the observer responded, whichever came first. Misses (i.e., failure to respond within 60 sec) were more common for the more difficult items, such that observers who missed more scenes would have artificially lower mean response times (RTs), given that the longest RTs (those to the difficult items that were missed) would be absent from their distribution. To avoid this speed–accuracy trade-off, only 51 items with accuracy rates over 90% within each age group were analyzed.

Of primary interest was the interaction of age with two item characteristics: the meaningfulness to driving of the change—that is, the extent to which the driver in the scene would need to pay attention to the changed object—and the salience of the change—that is, how visually conspicuous the change was within the scene. Item characteristics were obtained from a previous study in which independent observers rated each change on a scale of 0 to 5 for meaning

and for salience; ratings were then averaged to create one rating for each item (Pringle, Irwin, Kramer, & Atchley, 2001). Data were collected from 153 persons: 96 younger adults (41 men and 55 women, $M = 19.7$ years, $SD = 2.3$ years, range, 18–32) and 57 older adults (20 men and 37 women, $M = 75.7$ years, $SD = 5.4$ years, range, 63–86). The analysis was originally planned as a 2 (age group: young, old) × 2 (change meaning: low, high) × 2 (change salience: low, high) split-plot factorial ANOVA. Several issues would need to be addressed before proceeding with such an analysis, however.

Analytic treatment. The first issue has to do with the influence of accuracy on the available RTs. Although only scenes with accuracy levels over 90% were included, the data are still unbalanced because accuracy is not perfect, and the responses that are missing (because the change was not detected within 60 sec) are likely to be the responses to the most difficult items. Thus, the most difficult conditions, low meaning and low salience, are likely to have fewer responses contributing to the condition mean. As a result, those individual condition means may be less reliable or artificially improved (i.e., individual mean RTs would be too low because the items that would have had the highest RTs were not included), or may be missing entirely for some individuals, resulting in listwise deletion for those persons. Analyzing individual condition means without accounting for item missingness within the conditions will likely lead to biased estimates of the effects of the variables that are related to the probability of missingness (i.e., of nonresponse due to the imposed time limits in this case). A multilevel model would likely provide more accurate estimates in the presence of missing responses than would an ANOVA model; and, because listwise deletion would not be required, more observers could be included in the model, resulting in greater statistical power to detect the effects of interest.

The second issue is how to include the variable of age in the model. Although two distinct age groups were sampled, one ranging from 18 to 32 years and the other from 63 to 86 years, the older adults are likely to be considerably more heterogeneous in their RTs than the younger adults. Treating age as a dichotomous variable would therefore likely misrepresent the differences among older individuals varying in age, so that a 63-year-old might be expected to have the same score as an 86-year-old. Separating the older adults into two groups of "young-old" (under age 75) and "old-old" (age 75 or older), as is often done in experimental studies of aging, would also be inappropriate, because this assumes that a person of 74 is more like a person of 63 than like a person of 75. A multilevel model can allow a more accurate depiction of the effect of age on RTs as a semicontinuous (or *piecewise*) effect. The continuous age variable is therefore recoded into two variables: *old age*, in which persons 18 to 30 years old were coded as 0 and persons 65 and older were coded as 1; and *years over 65*, in which persons 18 to 30 years old were again coded as 0 but persons 65 and older were coded as their current age minus 65. Thus, the main effect of age on response time is represented with two piecewise slopes: (1) the slope of old age, representing the mean difference between the younger adults and 65-year-olds; and (2) the slope of years over 65, representing the additional increase in RT per year of age over 65. Additionally, because older adults are often more variable from one another than are younger adults (i.e., greater between-person variation), and also show more variability in their own responses across trials than do younger adults (i.e., greater within-person variation), separate random intercept and residual variances will be estimated for younger and older adults.

A similar problem concerns the distributions of change meaning and change salience across scenes. The assignment of items into low and high conditions for an ANOVA assumes bimodal distributions of change meaning and change salience, such that all items within each low or high condition are expected to have equivalent RTs. In this study, however, change meaning and change salience were measured in natural scenes, not manipulated, resulting in a continuous distribution for each. Thus, a median split would have been needed to create (artificial) categories of low and high, a practice with well-known problems of reduced power and increased Type I error, as discussed earlier. However, such distortion of the item-level or individual-level predictors is unnecessary in a multilevel model, in which categorical or continuous predictors can be easily accommodated at any level.

A multilevel analysis requires data to be structured differently than in repeated measures ANOVA, in which the data often need to be structured as *multivariate*, *wide*, or *person-level*, where each person's data is in a single row and the response variables per scene are in separate columns. In contrast, Table 2 provides an example of the data structure required for a multilevel analysis. In this structure, known as *stacked*, *long*, or *person-period*, each row contains the data for a single item for a single person. The current study has 7,803 rows of data, or 51 items multiplied by 153 persons. Variables relating to each person (e.g., ID, age) are copied down throughout the rows for each person, and variables relating to the items, such as change meaning, salience, and response time, are in each row. Item response times vary across subjects, but item characteristics are the same. SPSS and SAS syntax for combining multivariate data sets of subjects' responses and scene characteristics into a single stacked data set are available online (see Author Note).

Model specification. Five multilevel models were estimated using maximum likelihood (syntax available online; see the Author Note). The presence of incomplete data requires a choice in estimating denominator degrees of freedom, although differences among methods are likely to be trivial, except with small sample sizes. We used a commonly implemented strategy, the Satterthwaite method (see Fitzmaurice et al., 2004). Model 1 is an intercept-only or empty model, to be used as a baseline with which to assess the fit of more complex models, as given in Equation 5:

$$\begin{aligned} \text{Level 1:} \quad & y_{ti} = \beta_{0i} + W_t + e_{ti} \\ \text{Level 2:} \quad & \beta_{0i} = \gamma_{00} + U_{0i}, \end{aligned} \tag{5}$$

where y_{ti} is the natural log of RT in seconds of individual i and item t. RT was natural-log-transformed to reduce skew-

Table 2
Age Group Stacked Data Structure for Three Individuals and Five Items

Individual-Level Variables				Item-Level Variables					
ID	Age	OldAge	Yrs65	Item	Meaning	Salience	C_Mean	C_Sal	LN_RT
1	20	0	0	1	1	1	−2	−2	2.44
1	20	0	0	2	1	3	−2	0	2.37
1	20	0	0	3	3	3	0	0	2.29
1	20	0	0	4	3	1	0	−2	2.21
1	20	0	0	5	4	4	1	1	2.13
2	65	1	0	1	1	1	−2	−2	2.21
2	65	1	0	2	1	3	−2	0	2.13
2	65	1	0	3	3	3	0	0	2.06
2	65	1	0	4	3	1	0	−2	1.98
2	65	1	0	5	4	4	1	1	1.90
3	80	1	15	1	1	1	−2	−2	1.78
3	80	1	15	2	1	3	−2	0	1.64
3	80	1	15	3	3	3	0	0	1.50
3	80	1	15	4	3	1	0	−2	1.36
3	80	1	15	5	4	4	1	1	1.22

ness and to prevent spurious interactions with age due to baseline differences between younger and older adults (see Faust, Balota, Spieler, & Ferraro, 1999). In these equations, γs are used for fixed effects, U_is are used for individual random effects, and W_t is used for the random item effect. In the Level 1 model, β_{0i} is the intercept for individual i, derived from the following two parameters in the Level 2 model: the fixed intercept γ_{00}, or the grand mean across individuals and items; and the random intercept U_{0i}, or the individual-specific expected deviation about the grand mean. Finally, e_{ti} is the prediction error (Level 1 residual) for individual i and item t, or the difference between the observed and expected y_{ti} after accounting for individual i and item t. Thus, the variance of y is partitioned into three sources: the Level 2 between-subjects random intercept variance, which can be accounted for by subject-level variables such as age; the Level 2 between-items variance (i.e., random item variance), which can be accounted for by item-level variables such as change meaning and change salience; and the Level 1 trial-to-trial residual variance, which could be accounted for by trial-specific variables (e.g., order), but which will remain unaccounted for in this example.

Model 2A is a main effects model with homogeneous variances, as given in Equation 6:

$$\begin{aligned}
\text{Level 1:}\quad y_{ti} &= \beta_{0i} + \beta_{1i}\left(\text{Mean}_{ti} - 3\right) \\
&\quad + \beta_{2i}\left(\text{Sal}_{ti} - 3\right) + W_t + e_{ti} \\
\text{Level 2:}\quad \beta_{0i} &= \gamma_{00} + \gamma_{01}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{02}\left(\text{Years over 65}_i\right) + U_{0i}, \\
\beta_{1i} &= \gamma_{10},\ \ \beta_{2i} = \gamma_{20}, \qquad (6)
\end{aligned}$$

where γ_{00}, γ_{10}, and γ_{20} represent the fixed (main) effects of the intercept, meaning, and salience, respectively; γ_{01} and γ_{02} represent the fixed (main) effects of old age and years over 65 on the intercept, respectively. Meaning and salience were each centered at 3 (range, 0–5). Note that the interpretation of the fixed effect intercept γ_{00} has shifted, given that the intercept represents the expected value of y_{ti} when all other terms equal 0. Thus, γ_{00} now represents the expected RT for a younger adult (old age = 0; years over 65 = 0) for an item with meaning = 3 and salience = 3 (centered meaning = 0; centered salience = 0). The individual intercept β_{0i} is now a function of the fixed intercept γ_{00}, the fixed slope for old age γ_{01}, the fixed slope for years over 65 γ_{02}, and the random intercept U_{0i} representing the individual intercept deviation after controlling for age. Individual random effects were included for the intercept only. This assumption of only one source of individual differences (i.e., in the intercept) is a useful starting point, as estimation becomes considerably more difficult with multiple random effects. However, individual random effects for meaning and salience were examined in preliminary analyses and did not contribute significantly to the model, which suggests that these effects should be fixed.

Model 2A assumes that the magnitude of each component of variance is comparable across younger and older adults. However, it is reasonable that the sample of older adults will show greater variability than the sample of younger adults, both between subjects and across trials. The tenability of this assumption is tested in Model 2B, as seen in Equation 7:

$$\begin{aligned}
\text{Level 1:}\quad y_{ti} &= \beta_{0i} + \beta_{1i}\left(\text{Mean}_{ti} - 3\right) \\
&\quad + \beta_{2i}\left(\text{Sal}_{ti} - 3\right) + W_t + e_{ti}(\text{Y}) + e_{ti}(\text{O}) \\
\text{Level 2:}\quad \beta_{0i} &= \gamma_{00} + \gamma_{01}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{02}\left(\text{Years over 65}_i\right) \\
&\quad + U_{0i}(\text{Y}) + U_{0i}(\text{O}), \\
\beta_{1i} &= \gamma_{10},\ \ \beta_{2i} = \gamma_{20}, \qquad (7)
\end{aligned}$$

where Y is a dummy variable for old age = 0, and O is a dummy variable for old age = 1. Thus, although the fixed part is the same as in Model 2A, the error part of Model 2B now includes separate Level 1 (residual) and Level 2 (random intercept) variances for each age group.

Model 3A includes all two- and three-way interactions among meaning, salience, and old age, and among meaning, salience, and years over 65, as given in Equation 8:

$$
\begin{aligned}
\text{Level 1:}\quad y_{ti} &= \beta_{0i} + \beta_{1i}\left(\text{Mean}_{ti} - 3\right) \\
&\quad + \beta_{2i}\left(\text{Sal}_{ti} - 3\right) + \beta_{3i}\left(\text{Mean}_{ti} - 3\right) \\
&\quad \cdot \left(\text{Sal}_{ti} - 3\right) + W_t + e_{ti}(\text{Y}) + e_{ti}(\text{O}) \\
\text{Level 2:}\quad \beta_{0i} &= \gamma_{00} + \gamma_{01}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{02}\left(\text{Years over 65}_i\right) \\
&\quad + U_{0i}(\text{Y}) + U_{0i}(\text{O}), \\
\beta_{1i} &= \gamma_{10} + \gamma_{11}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{12}\left(\text{Years over 65}_i\right), \\
\beta_{2i} &= \gamma_{20} + \gamma_{21}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{22}\left(\text{Years over 65}_i\right), \\
\beta_{3i} &= \gamma_{30} + \gamma_{31}\left(\text{Old Age}_i\right) \\
&\quad + \gamma_{32}\left(\text{Years over 65}_i\right), \qquad (8)
\end{aligned}
$$

where model parameters are the same as in the main effects Model 2B, although they are conditional on the higher order interactions that have now been added: γ_{30} represents the fixed effect of the two-way interaction of meaning × salience; γ_{11}, γ_{21}, and γ_{31} represent the fixed effects of the two-way interactions of old age × meaning, old age × salience, and the three-way interaction of old age × meaning × salience, respectively; and γ_{12}, γ_{22}, and γ_{32} represent the fixed effects of the two-way interactions of years over 65 × meaning, years over 65 × salience, and the three-way interaction of years over 65 × meaning × salience, respectively. Thus, each individual slope for meaning, salience, and the two-way interaction of meaning × salience (β_{1i}, β_{2i}, and β_{3i}, respectively) depends on the fixed effect for the sample and the individual's values of old age and years over 65. A restricted version of Model 3A will also be estimated without any nonsignificant interactions (Model 3B).

Results

In the empty Model 1, the fixed intercept was 1.62, the expected natural-log-transformed RT in seconds for an average individual on an average item (i.e., the grand mean). The random intercept variance was 0.18, which represents the magnitude of the differences in overall RT across individuals. The random intercept variance can be interpreted in a standard deviation metric within a confidence interval, such that 95% of the sample would be expected to have an individual intercept between 0.77 and 2.47 ($1.62 \pm 2\sqrt{0.18}$), assuming an average item. The random item variance was 0.12, such that 95% of the items would be expected to have an intercept between 0.93 and 2.31, assuming an average individual. The residual variance is 0.39, the trial-to-trial variance in RT not accounted for by individuals or items. Thus, of the total variance (0.69), 26% is between subjects, 17% is between items, and 57% is between trials (i.e., an item by individual interaction; see also Raudenbush & Bryk, 2002).

Model 2A included main effects of meaning, salience, old age, and years over 65, each of which was significant. As seen in Table 3, the fixed effects for meaning (− 0.05) and salience (− 0.13) represent the expected linear rate of decline in response time for a one-unit increase in meaning or salience, respectively. The fixed effects for old age (0.59) and years over 65 (0.02) represent the expected difference in RT between younger adults and adults age 65 and the expected linear rate of increase in RT per year over 65, respectively.

In addition to significance tests for the fixed effects, however, the overall model −2 log likelihood value, or deviance, can be used to assess improvements in model

Table 3
Multilevel Model Parameters From Example 1

	Model 2A		Model 2B		Model 3B	
Parameter	Est	*SE*	Est	*SE*	Est	*SE*
Fixed Effects						
Intercept (γ_{00})	1.307***	0.045	1.306***	0.044	1.308***	0.046
Meaning (γ_{10})	−0.052*	0.023	−0.055*	0.023	−0.064*	0.023
Salience (γ_{20})	−0.132***	0.040	−0.134***	0.040	−0.143***	0.041
Old age (γ_{01})	0.590***	0.055	0.590***	0.070	0.614***	0.070
Years over 65 (γ_{02})	0.020***	0.004	0.020***	0.006	0.020***	0.006
Meaning × salience (γ_{30})					−0.003	0.019
Meaning × old age (γ_{11})					0.038***	0.009
Salience × old age (γ_{21})					0.013	0.015
Meaning × salience × old age (γ_{31})					−0.025***	0.007
Variance Components†						
Random item variance (W_t)	0.087***	0.018	0.088***	0.018	0.088***	0.018
Random intercept variance (U_{0i})	0.023***	0.004	0.011***	0.002	0.011***	0.003
			0.043***	0.010	0.043***	0.010
Residual variance (e_{ti})	0.390***	0.007	0.324***	0.007	0.323***	0.007
			0.507***	0.014	0.502***	0.014
Fit Statistics						
ML deviance (number of parameters)	14,885 (8)		14,692 (10)		14,657 (14)	
AIC; BIC	14,901; 14,917		14,712; 14,732		14,687; 14,712	

Note—AIC, Akaike information criterion; BIC, Bayesian information criterion. *$p < .05$. ***$p < .001$. †First value = younger; second value = older when two values are given.

fit. However, the models to be compared must include the exact same cases for the model deviance values to be comparable. The difference between two nested models in their deviance values is chi-square distributed as a function of the difference in the number of parameters estimated. Models that differ in fixed or random effects must be compared under maximum likelihood instead of restricted maximum likelihood, which is used for comparing models that differ in random effects or error structures only. Because we wanted to compare models differing in fixed effects, maximum likelihood was used to estimate each model. In addition, the AIC and BIC statistics also assess model fit relative to degrees of freedom, such that smaller values indicate a relatively better model. See Singer and Willett (2003) or Snijders and Bosker (1999) for more information about assessing model fit.

A comparison of model deviances suggested that main effects Model 2A was a significant improvement over the empty Model 1 [χ^2 difference (4) = 293, $p < .001$] and had smaller AIC and BIC values as well. Heterogeneity of variance across age groups was then examined in Model 2B. By comparing model deviances, it appears that the heterogeneous errors Model 2B was a significant improvement over homogeneous errors Model 2A [χ^2 difference (2) = 193, $p < .001$] and had smaller AIC and BIC values as well. As shown in Table 3, younger adults had significantly less between-subjects variation and less trial-to-trial variability as well.

The interaction Model 3A was then estimated (i.e., all two- and three-way fixed effect interactions among meaning, salience, and old age, and among meaning, salience, and years over 65). Although it was a significant improvement over Model 2B [χ^2 difference (7) = 40, $p < .001$] all of the interaction terms were nonsignificant. As such, beginning with the highest order, interaction terms were removed separately in sequential models in order to improve the parsimony of the overall model. The revised Model 3B (as seen in Table 3) did not include any interactions with years over 65, and was still a significant improvement over Model 2B [χ^2 difference (4) = 35, $p < .001$] and had smaller AIC and BIC values than Model 2B as well. All of the main effects and the interaction of meaning × old age were significant. Although the interactions of meaning × salience and salience × old age were not significant, the three-way interaction of meaning × salience × old age was significant. Figure 1 displays the expected fixed effects of salience at levels of low (1) and high (4) meaning for a younger adult, a person of 65, and a person of 80. As shown, RT increased with age and decreased with salience. For younger adults, RT decreased with meaning equivalently across levels of salience. For all older adults, however, the effect of meaning increased with salience.

Discussion

Example 1 used a crossed random effects multilevel model to examine the effects of between-subjects predictors (age) and between-item predictors (change meaning and salience) on RT in a change detection task. Because the multilevel model does not require listwise deletion for missing responses, using instead full-information maximum likelihood to estimate parameters on the basis of all available data, the multilevel model is likely to be more powerful than repeated measures ANOVA. The multilevel model also offers greater flexibility in examining the effects of categorical, semicontinuous, or continuous predictors at each level of analysis, as well as in allowing between-person and residual variances of different magnitudes across groups.

Example 2: Multivariate Analysis of RT and Error Rate

Research design. The second example was taken from part of a larger study (Hoffman, 2004) that used a visual search task to examine the effects of age and number of distractors on target detection time and error rate. Observ-

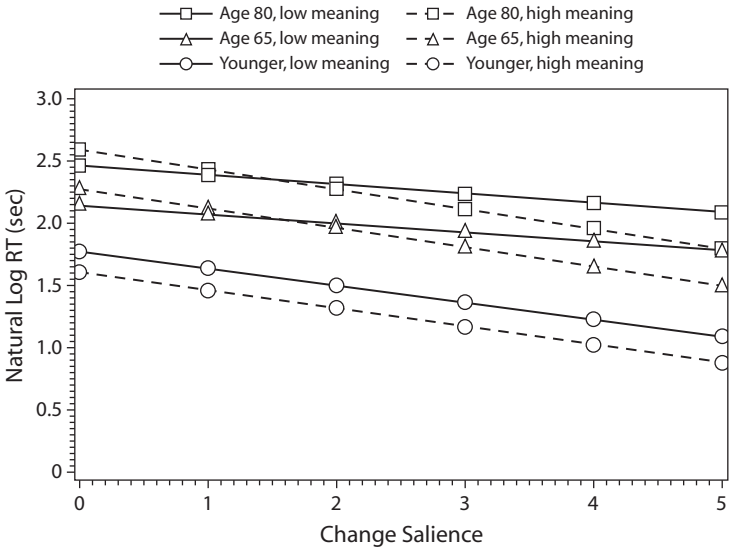

Figure 1. Results from Example 1. Expected effects of salience at low (1) and high (4) levels of meaning for an 80-year-old, 65-year-old, and younger adult.

ers searched for either an "L" or an "R" in a circular display of 3, 6, or 9 distractor letters. The initial task display had a fixation cross in the center surrounded by a black ring with a diameter of 3º visual angle presented for 750 msec, followed by 3, 6, or 9 black capital letters in 16-point bold font displayed for 294 msec. Each letter randomly occupied 1 of 18 places around the ring, with no two adjacent positions occupied. Participants responded to the "L" or the "R" by pressing a key with their left or right hand, respectively, within 5 sec. After practicing the task, 15 trials per target and set size were completed in a random order by 148 older adults (63 men, 85 women, $M = 75.3$ years, $SD = 4.7$ years; range, 63–87).

Analytic treatment. The analysis was envisioned as a 2 (target letter) × 3 (set size) repeated measures ANOVA with the effect of age as a covariate (i.e., ANCOVA). In Example 1, the units at the within-subjects level consisted of digital photographs with design factors measured along two continuous dimensions, which could, however, differ considerably in unmeasured dimensions. Conversely, in the present example, the design factors that differentiated the trials were manipulated by the experimenter, and thus trials of the same type (letter × set size) were expected to differ only slightly in their RTs. Given that the effects of target letter and increasing numbers of distractors could be seen through increased error rates as well as through increased RTs, however, it is important to consider both as indicators of performance. Only responses for correct trials were included; therefore, RTs and errors could not be modeled simultaneously at the trial level. The mean RT and error rate of the 15 trials in each condition were therefore modeled instead, as is typical in experimental studies. In contrast to typical analyses in experimental studies, however, RTs and error rates were modeled simultaneously in a multivariate model for the 6 conditions administered to each of the 148 subjects, rather than in separate univariate analyses. Conditions were treated as nested within subjects, given that the specific trials with correct RTs that were included in the condition means varied across subjects. Syntax for transforming the multivariate data set into a stacked data set for a multivariate analysis is available online (see Author Note).

Because error rate was the only source of missing data and was explicitly included in the model as a second outcome, any negative bias in the individual condition mean RTs across trials due to missing data (i.e., the noninclusion of incorrect trials in a more difficult condition) should be reflected in higher error rates for that condition. To that end, a multivariate model of RTs and error rates will be useful in evaluating a common concern in experimental studies, the speed–accuracy trade-off: the possibility that observers will slow down in order to preserve accuracy. It is important to note that, although multivariate analyses are possible within a repeated measures framework (e.g., by including a multivariate ANOVA test for a particular orthogonal trend across all dependent variables), separate analyses for each outcome are usually conducted instead. Further, although speed–accuracy trade-offs are thought to operate at the individual level, they are usually examined at the level of the aggregate sample. Mean differences in opposite directions for RT than for error rates are often taken as evidence of a speed–accuracy trade-off, the existence of which at the individual level—as is of primary interest—cannot be evaluated.

In the multivariate model, however, speed–accuracy trade-offs in terms of a correlation between mean RTs and error rates can be examined both within-subjects and between subjects. A negative within subjects correlation indicates that, within an individual, conditions that have lower RTs relative to the individual's RTs in other conditions are more likely to have relatively higher error rates. In contrast, a negative between-subjects correlation indicates that, if an individual has a lower overall RT relative to the rest of the sample he or she is also likely to have a relatively higher overall error rate. The consideration of both levels of analysis is likely to provide a more complete picture of speed–accuracy trade-offs than simply examining condition mean differences in the aggregate sample.

The multivariate model also permits comparisons of the magnitude of predictor effects across outcomes, provided that the outcomes are on the same metric. For example, the extent to which target letter and set size have greater effects on RTs than on error rates will be examined after transforming each outcome separately onto a unit-normal metric (i.e., *z*-score). Finally, although continuous age could be included as a main effect in a repeated measures analysis, its interaction with other predictors is much easier to examine in a multilevel model.

Model specification. Five multilevel models were estimated (SAS and SPSS syntax available online; see the Author Note). Model 1 is an intercept-only or empty multivariate model, to be used as a baseline with which to assess the fit of more complex models, as given in Equation 9:

$$\begin{aligned}
\text{Level 1:}\quad y_{tik} &= \beta_{0i1}(\text{DV1}) + e_{ti1}(\text{DV1}) \\
&\quad + \beta_{0i2}(\text{DV2}) + e_{ti2}(\text{DV2}) \\
\text{Level 2:}\quad \beta_{0i1} &= \gamma_{001} + U_{0i1}, \\
\beta_{0i2} &= \gamma_{002} + U_{0i2},
\end{aligned} \tag{9}$$

where y_{tik} and e_{tik} are the observed and residual values for condition t, individual i, and outcome k, where $k = 1$ indicates natural log response time in milliseconds, and $k = 2$ indicates proportion errors. DV1 and DV2 are dummy variables for each outcome. DV1 = 1 for RT and 0 for error rate, and DV2 = 0 for RT and 1 for error rate. The inclusion of the DV1 and DV2 dummy variables serves as a programming trick with which to obtain separate parameter estimates for the effects of the independent variables for each outcome. To illustrate, the expected values for each outcome are written out in Equation 10:

$$\begin{aligned}
\text{RT:}\quad y_{tik} &= \beta_{0i1}(1) + e_{ti1}(1) \\
&\quad + \beta_{0i2}(0) + e_{ti2}(0) \\
\text{Error Rate:}\quad y_{ti2} &= \beta_{0i1}(0) + e_{ti1}(0) \\
&\quad + \beta_{0i2}(1) + e_{ti2}(1),
\end{aligned} \tag{10}$$

where the terms not pertaining to each outcome (i.e., when $k = 2$ for RT, or $k = 1$ for error rate) are reduced to zero when multiplied by DV1 for error rate, or DV2 for RT.

Returning to Equation 9, β_{0i1} and β_{0i2} are the individual intercepts for RT and error rate, respectively, as derived from the fixed intercepts (i.e., grand means) for RT, γ_{001}, and error rate, γ_{002}, and the random intercept for individual i for response time, U_{0i1}, and error rate, U_{0i2}. The variance in each outcome is thus partitioned into between-subjects random intercept variance (the U_{0i}s) and within-subjects residual variance (the e_{ti}s). By estimating unconstrained matrices for the random effects and residual variances (G and R, respectively), each variance component is permitted to correlate across outcomes, as shown in Equation 11:

$$\mathrm{G} = \begin{bmatrix} v_{U_{01}} & \\ r_{21} & v_{U_{02}} \end{bmatrix} \quad \mathrm{R} = \begin{bmatrix} v_{e_1} & \\ r_{21} & v_{e_2} \end{bmatrix}. \tag{11}$$

Model 2 is a main effects only model, as given in Equation 12:

$$\begin{aligned}
\text{Level 1:}\quad y_{tik} &= \beta_{0i1}(\mathrm{DV1}) + \beta_{1i1}\left(\mathrm{S}_{ti} - 6\right)(\mathrm{DV1}) \\
&\quad + \beta_{2i1}\left(\mathrm{T}_{ti}\right)(\mathrm{DV1}) + e_{ti1}(\mathrm{DV1}) \\
&\quad + \beta_{0i2}(\mathrm{DV2}) + \beta_{1i2}\left(\mathrm{S}_{ti} - 6\right)(\mathrm{DV2}) \\
&\quad + \beta_{2i2}\left(\mathrm{T}_{ti}\right)(\mathrm{DV2}) + e_{ti2}(\mathrm{DV2}) \\
\text{Level 2:}\quad \beta_{0i1} &= \gamma_{001} + \gamma_{011}\left(\mathrm{age}_i - 75\right) + U_{0i1}, \\
\beta_{0i2} &= \gamma_{002} + \gamma_{012}\left(\mathrm{age}_i - 75\right) + U_{0i2}, \\
\beta_{1i1} &= \gamma_{101} + U_{1i1}, \\
\beta_{1i2} &= \gamma_{102} + U_{1i2}, \\
\beta_{2i1} &= \gamma_{201} + U_{2i1}, \\
\beta_{2i2} &= \gamma_{202} + U_{2i2},
\end{aligned} \tag{12}$$

where γ_{001}, γ_{101}, and γ_{201} represent the fixed (main) effects for response time of the intercept, set size, and target, respectively, and γ_{011} represents the fixed (main) effect of age on the intercept. The γs with $k = 2$ as a subscript represent the same parameters for error rates. Set size was centered at 6 and age was centered at 75 years, such that the fixed intercepts γ_{001} and γ_{002} now represent the expected RT and error rate, respectively, for a 75-year-old for Set Size 6, Target L. The individual intercepts for RT and error rate, β_{0i1} and β_{0i2}, are now a function of the fixed effect intercept for each outcome, γ_{001} and γ_{002}, the fixed (main) effect for age for each outcome, γ_{011} and γ_{012}, and the random intercepts for each outcome, U_{011} and U_{012}, which represent the individual's systematic deviation from the expected fixed intercepts after controlling for age. The individual effects of set size for each outcome, β_{1i1} and β_{1i2}, are derived from the fixed (main) effects of set size γ_{101} and γ_{201} and the random effects of set size U_{1i1} and U_{1i2}, which represent the individual's systematic deviation from expected effect of set size. The individual effects of target for each outcome, β_{2i1} and β_{2i2}, are similarly derived from the fixed (main) effects of target γ_{101} and γ_{201}, and the random effects of target U_{2i1} and U_{2i2}. Thus, in Model 2, the variance in RT and error rate is partitioned into four components: three between-subjects variances of the individual intercepts, slopes for set size, and slopes for target, and one within-subjects residual variance. Each variance component is again permitted to correlate across outcomes, as shown in Equation 13:

$$\mathrm{G} = \begin{bmatrix} v_{U_{01}} & & & & & \\ r_{21} & v_{U_{02}} & & & & \\ r_{31} & r_{32} & v_{U_{11}} & & & \\ r_{41} & r_{42} & r_{43} & v_{U_{12}} & & \\ r_{51} & r_{52} & r_{53} & r_{54} & v_{U_{21}} & \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & v_{U_{22}} \end{bmatrix}$$

$$\mathrm{R} = \begin{bmatrix} v_{e_1} & \\ r_{21} & v_{e_2} \end{bmatrix}. \tag{13}$$

Model 3A includes all interactions among set size, target, and age, as seen in Equation 14:

$$\begin{aligned}
\text{Level 1:}\quad y_{tik} &= \beta_{0i1}(\mathrm{DV1}) + \beta_{1i1}\left(\mathrm{S}_{ti} - 6\right)(\mathrm{DV1}) \\
&\quad + \beta_{2i1}\left(\mathrm{T}_{ti}\right)(\mathrm{DV1}) \\
&\quad + \beta_{3i1}\left(\mathrm{S}_{ti} - 6\right)\left(\mathrm{T}_{ti}\right)(\mathrm{DV1}) \\
&\quad + e_{ti1}(\mathrm{DV1}) \\
&\quad + \beta_{0i2}(\mathrm{DV2}) + \beta_{1i2}\left(\mathrm{S}_{ti} - 6\right)(\mathrm{DV2}) \\
&\quad + \beta_{2i2}\left(\mathrm{T}_{ti}\right)(\mathrm{DV2}) \\
&\quad + \beta_{3i2}\left(\mathrm{S}_{ti} - 6\right)\left(\mathrm{T}_{ti}\right)(\mathrm{DV2}) \\
&\quad + e_{ti2}(\mathrm{DV2}) \\
\text{Level 2:}\quad \beta_{0i1} &= \gamma_{001} + \gamma_{011}\left(\mathrm{age}_i - 75\right) + U_{0i1}, \\
\beta_{0i2} &= \gamma_{002} + \gamma_{012}\left(\mathrm{age}_i - 75\right) + U_{0i2}, \\
\beta_{1i1} &= \gamma_{101} + \gamma_{111}\left(\mathrm{age}_i - 75\right) + U_{1i1}, \\
\beta_{1i2} &= \gamma_{102} + \gamma_{112}\left(\mathrm{age}_i - 75\right) + U_{1i2}, \\
\beta_{2i1} &= \gamma_{201} + \gamma_{211}\left(\mathrm{age}_i - 75\right) + U_{2i1}, \\
\beta_{2i2} &= \gamma_{202} + \gamma_{212}\left(\mathrm{age}_i - 75\right) + U_{2i2}, \\
\beta_{3i1} &= \gamma_{301} + \gamma_{311}\left(\mathrm{age}_i - 75\right), \\
\beta_{3i2} &= \gamma_{302} + \gamma_{312}\left(\mathrm{age}_i - 75\right),
\end{aligned} \tag{14}$$

where model parameters are the same as in the main effects Model 2, although they are conditional on the higher order interactions that have now been added: γ_{301} represents the fixed effect for RT of the interaction of set size × target; γ_{111}, γ_{211}, and γ_{311} represent the fixed effects for RT of the interactions of age × set size, age × target, and the three-way interaction of age × set size × target, respectively. The γs with $k = 2$ as a subscript represent the same parameters for error rates. Thus the individual slopes of set size and target for each outcome depend on the fixed effect, the random effect, and the individual's value of age. The individual slopes of the interaction of set size and target for each outcome depend only on the fixed effect and

the individual's value of age (random effects were again not included for the interaction). A restricted version of Model 3A will also be estimated without any nonsignificant interactions (Model 3B).

Finally, the multivariate model can be reparameterized into Model 4 in order to examine whether each fixed effect is of different magnitude across outcomes, as shown in Equation 15:

$$
\begin{aligned}
\text{Level 1:}\quad y_{tik} &= \beta_{0i1} + \beta_{1i1}\left(\mathrm{S}_{ti} - 6\right) + \beta_{2i1}\left(\mathrm{T}_{ti}\right) \\
&\quad + \beta_{3i1}\left(\mathrm{S}_{ti} - 6\right)\left(\mathrm{T}_{ti}\right) + e_{ti1}(\mathrm{DV1}) \\
&\quad + \beta_{0i2}(\mathrm{DV2}) + \beta_{1i2}\left(\mathrm{S}_{ti} - 6\right)(\mathrm{DV2}) \\
&\quad + \beta_{2i2}\left(\mathrm{T}_{ti}\right)(\mathrm{DV2}) \\
&\quad + \beta_{3i2}\left(\mathrm{S}_{ti} - 6\right)\left(\mathrm{T}_{ti}\right)(\mathrm{DV2}) \\
&\quad + e_{ti2}(\mathrm{DV2}) \\
\text{Level 2:}\quad \beta_{0i1} &= \gamma_{001} + \gamma_{011}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{0i1}(\mathrm{DV1}), \\
\beta_{0i2} &= \gamma_{002} + \gamma_{012}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{0i2}(\mathrm{DV2}), \\
\beta_{1i1} &= \gamma_{101} + \gamma_{111}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{1i1}(\mathrm{DV1}), \\
\beta_{1i2} &= \gamma_{102} + \gamma_{112}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{1i2}(\mathrm{DV2}), \\
\beta_{2i1} &= \gamma_{201} + \gamma_{211}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{2i1}(\mathrm{DV1}), \\
\beta_{2i2} &= \gamma_{202} + \gamma_{212}\left(\mathrm{age}_i - 75\right) \\
&\quad + U_{2i2}(\mathrm{DV2}), \\
\beta_{3i1} &= \gamma_{301} + \gamma_{311}\left(\mathrm{age}_i - 75\right), \\
\beta_{3i2} &= \gamma_{302} + \gamma_{312}\left(\mathrm{age}_i - 75\right), \qquad (15)
\end{aligned}
$$

where the DV1 dummy variable is no longer included in the fixed effects (although it remains in the random effects and residual errors so that separate variance components are estimated for each outcome), and there is now only one true fixed intercept. Although statistically equivalent to Model 3A, this model parameterization allows for tests of the differences in the magnitude of the fixed effects across outcomes. Specifically, the γs with $k = 1$ represent the same parameters as before (i.e., fixed effects for RT), whereas the γs with $k = 2$ now represent the *difference in the fixed effects between outcomes*. For example, a significant γ_{012} parameter would indicate that the main effect of age is different for RT than for error rate.

Recall that because both outcomes must be on the same metric in order for this specification to be meaningful, RT and error rate were thus each transformed onto a unit-normal (z-score) metric prior to estimating Model 4, so that all parameter estimates refer to standard deviation units (i.e., standardized coefficients). This transformation does remove any differences in the magnitude of variability across outcomes, however. If one is interested in differential magnitudes of variability across response variables on different metrics, then multivariate tests cannot be used as described here.

Results

Table 4 provides the parameter estimates and fit statistics from each model. Model 1 is an empty baseline model. The fixed RT intercept was 6.73 (95% CI = 6.37 to 7.01), and for error rate was .17 (95% CI = 0.02 to 0.33), which are the expected natural-log-transformed RT in milliseconds and proportion error rate for any individual for any condition (i.e., the grand means), respectively. The intraclass correlations for RT and error rate, calculated by dividing the random intercept variance by the total variance (Snijders & Bosker, 1999), were .57 and .34, indicating that 57% and 34% of the variance in RT and error rate was between subjects and 43% and 66% was within subjects, respectively. Model 1 also provides unconditional covariances between RT and error rate at the between- and within-subjects levels (i.e., before controlling for any predictors), from which correlations may be calculated (covariance / [SQRT(var1) * SQRT(var2)]). Although the between-subjects or random intercept covariance was not significant ($r = .05$, $p > .05$), the within-subjects or residual covariance was significant ($r = .42$, $p < .001$), indicating that within individuals, conditions with higher response times also had higher error rates.

Model 2 included main effects of set size, target, and age, each of which was significant, as shown in Table 4. The fixed effects of set size for response time (.04, random effects 95% CI = .02 to .06) and error rate (.03, random effects 95% CI = .01 to .05) represent the expected linear rate of increase in each outcome per additional distractor. The confidence intervals for the random effects around the fixed effect of set size indicate that most individuals were predicted to experience greater RTs and error rates with increasing set size, with the rate of increase varying across individuals. The fixed effects of target for RT (.08, random effects 95% CI = −.11 to .27) and error rate (.04, random effects 95% CI = −.09 to .17) represent the expected difference in each outcome between the conditions, with the target R instead of L. The confidence intervals for the random effects around the effect of target indicate that not all individuals were predicted to experience greater RTs and error rates for Target R than for L, although this was true on average, as indicated by the direction of the fixed effect. The fixed effects of age for RT (.007) and error rate (.004) represent the expected linear rate of increase in each outcome per additional year of age. A comparison of Model 2 to a version with random effects for the intercept only revealed a significant decrease in fit [χ^2 difference (18) = 126, $p <$ .001], as well as larger AIC and BIC values, indicating that the effects of set size and target should be random, and thus do vary significantly over individuals.

It is important to note, however, that Model 2 assumes a linear effect of set size, in that only one slope for set size was specified. To test this assumption, a piecewise model specifying two fixed set size slopes (3–6 and 6–9)

Table 4
Response Time (RT) and Error Rate (ER) Multilevel Model Parameters From Example 2

	Model 1		Model 2		Model 3B	
Parameter	Est	*SE*	Est	*SE*	Est	*SE*
Fixed Effects						
RT intercept (γ_{001})	6.732***	0.016	6.690***	0.016	6.689***	0.016
RT set size (γ_{101})			0.036***	0.002	0.030***	0.002
RT target letter (γ_{201})			0.080***	0.010	0.080***	0.010
RT age (γ_{011})			0.007*	0.003	0.010*	0.003
RT set size by target letter (γ_{301})					0.012***	0.003
RT age by set size (γ_{111})					0.001*	0.000
ER intercept (γ_{002})	0.172***	0.008	0.152***	0.008	0.152***	0.008
ER set size (γ_{102})			0.027***	0.001	0.026***	0.001
ER target letter (γ_{202})			0.039***	0.008	0.039***	0.008
ER age (γ_{012})			0.004*	0.002	0.003*	0.001
Variance Components						
RT intercept variance (U_{0i1})	0.0331***	0.0043	0.0332***	0.0042	0.0333***	0.0042
RT set size variance (U_{1i1})			0.0001*	0.0003	0.0001*	0.0000
RT target letter variance (U_{2i1})			0.0095***	0.0019	0.0098***	0.0019
RT residual variance (e_{ti1})	0.0250***	0.0013	0.0098***	0.0007	0.0093***	0.0006
ER intercept variance (U_{0i2})	0.0060***	0.0010	0.0070***	0.0012	0.0069***	0.0012
ER set size variance (U_{1i2})			0.0001*	0.0000	0.0001*	0.0000
ER target letter variance (U_{2i2})			0.0042***	0.0012	0.0042***	0.0012
ER residual variance (e_{ti2})	0.0161***	0.0010	0.0088***	0.0006	0.0087***	0.0006
RT–ER intercept covariance	0.0007	0.0010	0.0024	0.0016	0.0026	0.0016
RT–ER set size covariance			0.0000	0.0000	0.0000	0.0000
RT–ER target letter covariance			0.0008	0.0011	0.0008	0.0011
RT–ER residual covariance	0.0084***	0.0010	0.0001	0.0004	0.0000	0.0004
Fit Statistics						
ML deviance (number of parameters)	−1,550 (8)		−2,308 (32)		−2,331 (34)	
AIC; BIC	−1,534; −1,510		−2,244; −2,148		−2,263; −2,160	

*$p < .05$. ***$p < .001$. Values of .0000 are 0 to the fourth decimal place, but are not exactly 0.

was compared to Model 2. Both models included random intercepts only, however, due to estimation problems with the random effects with the piecewise model. Although the piecewise model had marginally better fit than the linear model [χ^2 difference (2) = 6.8, $p = .04$], the BIC value favored the linear model. The linear model was retained on the basis of parsimony (i.e., to limit the number of parameters in estimating interactions with other variables) and in order to include random effects for set size and target.

The interaction Model 3A (all two- and three-way fixed effect interactions among set size, target, and age) was then estimated. Although it was a significant improvement over Model 2 [χ^2 difference (8) = 33, $p < .001$], only the interaction terms of set size × target and age × set size were significant for RT, and no interaction terms were significant for error rate. The nonsignificant interaction terms were then removed separately in sequential models in order to improve the parsimony of the model. The revised Model 3B still had significantly better fit than Model 2 [χ^2 difference (2) = 24, $p < .001$], and had smaller AIC and BIC values than Model 2 as well. All fixed effects were significant, as shown in Table 4. Figures 2A and 2B display the expected fixed effects of set size for each target letter for a 65-, 75-, and 85-year-old, for RT and error rate, respectively. RT increased with age and set size, and the effect of set size increased with age. RTs were higher to the Target R than L, and this difference increased with set size. Error rates also increased with age and set size, and error rates were higher when responding to a Target R than L, but no interactions were present.

Model 3B also provides correlations between RT and error rate at the between-subjects (random effects of intercept, set size, and target) and within-subjects (residual) levels, conditional on the effects of set size and target. The within-subjects covariance was no longer significant, indicating that there was no relationship between RT and error rate across conditions within individuals, after controlling for the effects of set size and target. Neither of the covariances between the random effects for set size and target was significant, indicating that individuals who displayed a larger effect of set size for RT, relative to the rest of the sample, did not necessarily display a relatively larger effect of set size for error rates, with a similar interpretation for the random effects of target. The covariance between the random intercepts between persons was marginally significant, however ($r = .17$, $p = .05$), indicating that individuals with higher *overall* RTs relative to the rest of the sample also had relatively higher *overall* error rates. This is the opposite of a speed–accuracy trade-off.

Finally, the extent to which the effects of set size, target, and age were different across outcomes was examined in Model 4 using the standardized response variables, although only the interactions of set size × target and age × set size were included based on previous results. The interaction with DV2 of set size was significant, indicating that the effect of set size on response time (.13 *SD*) was significantly smaller than the effect of set size on error rate (.17 *SD*). The interactions with DV2 of target, age, and set size × target were not significant, however, indicating the effects of target on RT (.33 *SD*) and error

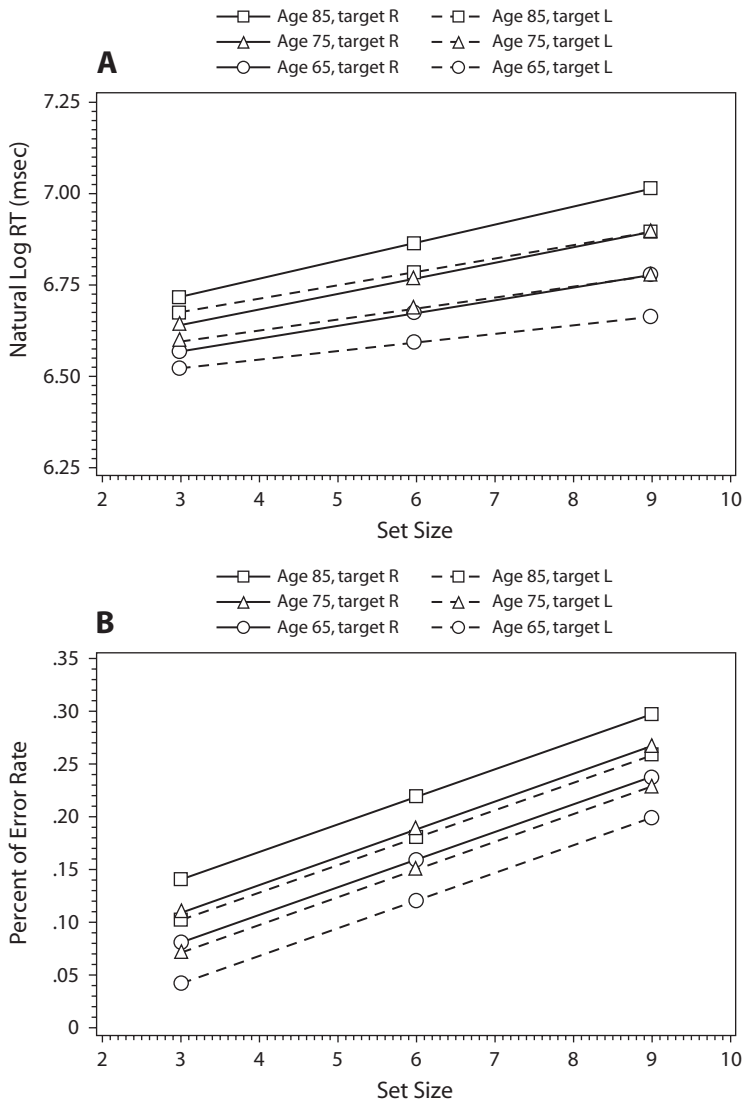

Figure 2. Results from Example 2. Expected effects of set size for target letter R and L for an 85-year-old, 75-year-old, and 65-year-old for response time (RT) (A) and error rate (B).

rate (.26 *SD*) were equivalent, as were the effects of age on RT (.04 *SD*) and error rate (.02 *SD*), as well as the effects of set size × target on RT (.05 *SD*) and error rate (.03 *SD*). The interaction with DV2 of age × set size was marginally significant ($p = .06$), such that the interaction of age × set size on response time (.003 *SD*, which was significant) was significantly larger than the interaction of age × set size on error rate (−.002 *SD*, which was not significant).

Discussion

Example 2 used a multivariate multilevel model to examine the effects of between-subjects variables (age) and within-subjects variables (set size, target letter) simultaneously on RT and error rate in a visual search task. In addition to the general advantages of the multilevel model discussed in Example 1 (e.g., inclusion of incomplete responses, categorical or continuous predictors at each level), the multivariate multilevel model can estimate correlations between outcomes at the within-subjects and between-subjects levels, and can also permit tests of differences in the magnitude of the predictor effects across outcomes. In Example 2, no evidence of a speed–accuracy trade-off was found—in fact, the correlations between RT and error rate were actually marginally positive instead of significantly negative—and effect sizes of the predictors were shown to be equivalent across outcomes, with the exception of the effect of set size (significantly smaller for RT) and the effect of age × set size (significantly larger for RT).

SUMMARY AND CONCLUSIONS

The purpose of this article was to illustrate how the multilevel or general linear mixed model can be used in

the analysis of data from experimental designs. The multilevel model is relatively common in the educational and developmental literatures, but is less well known in other areas of psychology, with a few exceptions (see Allen, Sliwinski, & Bowie, 2002; Quené & van den Bergh, 2004; Wright, 1998). Although the repeated measures ANOVA model has a well earned place in the toolbox of the experimental psychologist, there are many scenarios in which the assumptions of a repeated measures ANOVA may not be met, or the model may be too restrictive, and in which case a multilevel model might be more useful. These scenarios include: (1) main effects and interactions of continuous or semicontinuous predictors pertaining to experimental stimuli or individuals, (2) different magnitudes of between-subjects and within-subjects residual variances across groups, (3) violations of compound symmetry resulting from sources of variance related to individual differences, (4) the presence of nested observations or crossed random effects, (5) the presence of incomplete data that would require listwise deletion or otherwise result in bias and loss of power, and (6) the desire to examine differences in effect sizes and multivariate relations across outcomes at multiple levels of analysis. In presenting two in-depth examples from the experimental literature along with SAS and SPSS program syntax for data restructuring and analysis, we hope this article will be useful in providing guidance to investigators dealing with similar scenarios in the future.

AUTHOR NOTE

The SAS and SPSS syntax used to estimate the example models and the accompanying data sets are available electronically from the first author at psycweb.unl.edu/psypage/hoffman/HomePage.htm. We thank Deborah Eakin, Sarah Kollat, and Rob Stawski for their helpful comments in preparing this manuscript, and Devon Land, Alicia Mackay, Christian Stopp, and Imad Uddin for their assistance in collecting data. Correspondence concerning this article should be addressed to L. Hoffman, Department of Psychology, 238 Burnett Hall, University of Nebraska, Lincoln, NE 68588-0308 (e-mail: lhoffman2@unlnotes.unl.edu).

REFERENCES

Allen, P. A., Sliwinski, M. J., & Bowie, T. (2002). Differential age effects in semantic and episodic memory: Part II. Slope and intercept analyses. *Experimental Aging Research*, **28**, 111-142.

Allison, P. D. (1994). Using panel data to estimate the effects of events. *Sociological Methods & Research*, **23**, 174-199.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, **7**, 249-253.

Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, **21**, 171-192.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, **125**, 777-799.

Fitzmaurice, G., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: Wiley.

Hoffman, L. (2004). Attentional orienting deficits as a predictor of driving impairment in older adults. *Dissertation Abstracts International*, **65** (4), 2130B.

Hoffman, L., & Atchley, P. (2001, November). *Attentional orienting: The costs of age and the benefits of processing speed.* Poster presented at the 42nd Annual Meeting of the Psychonomic Society, Orlando, FL.

Huynh, H., & Feldt, L. S. (1980). Performance of traditional *F* tests in repeated measures designs under covariance heterogeneity. *Communications in Statistics: Theory & Methods*, **9**, 61-74.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, **30**, 1-21.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modeling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793-1819.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 149-157.

Maas, C. J. M., & Snijders, T. A. B. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality & Quantity: International Journal of Methodology*, **37**, 71-89.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, **7**, 19-40.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, **113**, 181-190.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, **97**, 316-333.

Pringle, H. L., Irwin, D. E., Kramer, A. F., & Atchley, P. (2001). The role of attentional breadth in perceptual change detection. *Psychonomic Bulletin & Review*, **8**, 89-95.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, **43**, 103-121.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with the "language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory & Language*, **41**, 416-426.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, **8**, 368-373.

Rovine, M. J., & von Eye, A. (1991). *Applied computational statistics in longitudinal research*. Boston: Academic Press.

Sayer, A. G., & Klute, M. M. (2004). Analyzing couples and families. In V. L. Bengtson, A. Acock, K. R. Allen, P. Dilworth-Anderson, & D. M. Klein (Eds.), *Sourcebook of family theory and research* (pp. 289-314). Thousand Oaks, CA: Sage.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147-177.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational & Behavioral Statistics*, **23**, 323-355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Wallace, D., & Green, S. B. (2002). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data* (pp. 103-134). Mahwah, NJ: Erlbaum.

Wright, D. B. (1998). Modelling clustered data in autobiographical memory research: The multilevel approach. *Applied Cognitive Psychology*, **12**, 339-357.

(Manuscript received October 17, 2005;
revision accepted for publication January 19, 2006.)