Analyzing individual differences in sentence processing performance using multilevel models

SHELLEY A. BLOZIS AND MATTHEW J. TRAXLER

University of California, Davis, California

The use of multilevel models is increasingly common in the behavioral sciences for analyzing hierarchically structured data, including repeated measures data. These models are flexible and easily implemented via a variety of commercially available statistical software programs. We consider their application in the context of an eye-movement experiment testing readers' responses to a semantic plausibility manipulation in temporarily ambiguous sentences. Multilevel models were used to study the relationship between working memory capacity and the extent to which readers were disrupted by syntactic misanalysis. This represented a cross-level interaction between an individual difference measure and a sentence-level characteristic. We compare a multilevel modeling approach to a standard approach based on ANOVA.

Psycholinguists have attempted to understand how individual differences in general cognitive abilities affect the processing strategies that comprehenders use to interpret language. In the sentence processing literature, much of the focus has been on the relationship between verbal working memory (WM) capacity and syntactic parsing (e.g., Baddeley, 1986; Baddeley & Hitch, 1974; Caplan & Waters, 1990, 1999; Just & Carpenter, 1992; Just, Carpenter, & Keller, 1996; King & Just, 1991; Pearlmutter & MacDonald, 1995; Traxler, Morris, & Seely, 2002; Traxler, Williams, Blozis, & Morris, 2005; Wanner & Maratsos, 1978; Waters & Caplan, 1996). Our main goal in this article is to review quasiexperimental data analysis strategies that have commonly been used to investigate individual differences in sentence processing, and to contrast them with an approach based on multilevel models that we believe offers significant advantages.

Traditional Analytic Approach to Individual Differences in Psycholinguistic Experiments

Typical studies of sentence processing require that subjects read multiple sentences in which some categorical text variable (such as syntactic ambiguity) is manipulated. Currently, experimental studies of sentence processing generally rely on aggregated data at the individual level. For each subject, responses are usually summed or averaged within levels of the text variable, with each subject contributing one datum per level. Standard analytic methods such as ANOVA may then be employed by treating the text factor as a within-subjects factor. Consequently, inference about subject performance is made at the subject level and not at the sentence level. Advantages of this approach are that it generally solves problems of missing data (assuming data are missing at random) in the event

that some subjects have incomplete data, and it simplifies the problem to a simple analytic design. Primary disadvantages of the approach are that inference about sentence processing is limited to the aggregated data level, and any features that distinguish one item from another are lost in the aggregation.

With regard to the study of individual differences in sentence processing, research has been hindered by problems related to the assessment of WM capacity (but see Waters & Caplan, 2003) and, more importantly, by the lack of sound statistical tools that can adequately evaluate data models. For example, in many studies assessing WM contributions to sentence processing performance, subjects are selected on the basis of scoring very high or very low on a WM test (most commonly the *sentence span* test, Daneman & Carpenter, 1980). Subjects then read sentences in which some categorical text variable (such as syntactic ambiguity) is manipulated. Responses within subjects are then aggregated within levels of the text variable. The data may then be analyzed using a design with one within-subjects factor and one between-subjects factor, with WM capacity (high vs. low) treated as the between-subjects factor. With the central distribution of the WM score excluded (by design), this type of analysis increases the risk of showing a spurious interaction of WM capacity and the text variable (e.g., Clifton et al., 2003; Pearlmutter & MacDonald, 1995). In other studies in which extreme-group, quasiexperimental design is not adopted, the conceptually continuous WM variable is nevertheless treated as a categorical variable—for example, by the creation of “high,” “medium,” and “low” capacity groups using arbitrary cutoff scores (this design is often used in other fields of cognitive research as well; see, e.g., Verschueren, Schaeken, & d’Ydewalle, 2005). In

addition to issues concerning the treatment of continuous individual difference measures, this approach, although easily implemented, ignores any possible interaction effects between subject-level variables, such as WM, and text characteristics. For example, certain text properties may differentially affect performance on the outcome measure, with such differences possibly accounted for, at least in part, by individual difference measures.

In this article, we contrast the traditional approach of using repeated measures ANOVA with a multilevel model (also known as a hierarchical linear model). A multilevel model is a regression-based approach that addresses hierarchical data structures, such as repeated measures. When repeated measures are taken, responses made by each subject are not, in most cases, independent. Ignoring dependence of responses within subjects can lead to biases in the estimates of the standard errors corresponding to fixed parameter estimates (see, e.g., Raudenbush & Bryk, 2002). Multilevel models address this dependency by partitioning variance in the responses into two levels: a within-subjects level and a between-subjects level. At the first level, the within-subjects errors address variation in responses after important predictor variables have been taken into account. At the second level, the effects (regression coefficients) of one or more of the predictors that enter the model at the first level are allowed to vary across subjects. These effects are commonly known as random coefficients or random effects. This means that the effect of a predictor at the first level can differentially affect the outcome, depending on the particular subject. The dependencies of scores within subjects are assumed to be accounted for by the Level 1 effects that vary across subjects (Snijders & Bosker, 1999).

Multilevel models offer a number of advantages over the traditional approach of using repeated measures ANOVA that are worthy of consideration in the context of sentence processing studies. First, responses at the sentence level are modeled directly in a multilevel model, thus eliminating the step of aggregating scores within subjects. Perhaps more importantly, this also allows one to directly account for characteristics that distinguish the individual sentences, such as sentence length, by including such variables in the model at the sentence level (Level 1). Subject-level variables enter at the second level of the model. Second, multilevel models allow for interactions between individual difference variables (e.g., WM capacity) and text-level variables (e.g., syntactic ambiguity), without treating the individual difference variables as categorical, without arbitrary divisions between groups of subjects, and without systematically excluding subsets of subjects based on near-average performance on the individual difference measure.

By specifying interactions between text-level variables and subject-level measures, it may be possible to evaluate potentially moderating effects of subject-level variables (possibly measured continuously) on the effects of text-level variables on the response. Such moderating effects are known as cross-level interactions, because they involve interactions between Level 1 variables (e.g., text-level variables) and Level 2 variables (subject-level variables).

Such interactions cannot be specified using standard approaches such as ANOVA. In the context of a sentence processing study, cross-level interactions might indicate individual differences in the effects of text variables on the outcome response, where such differences vary as functions of subject-level variables, such as WM capacity.

Finally, a multilevel model easily handles missing data for the outcome measure, because the estimation procedures typically used do not require that subjects have the same number of responses across or within conditions. When data are missing, estimation of the model may rely on full information maximum likelihood (FIML), an estimation procedure commonly employed in many software packages (see Snijders & Bosker, 1999, for a review). Analyses based on FIML use the raw data directly rather than relying on a mean vector or variance-covariance matrix estimated from the data. Oftentimes, missing data—such as those due to equipment failure or random errors committed by the respondent, including those due to fatigue or inattention—are assumed to be missing at random. General strategies for dealing with responses not missing at random are discussed in Little (1995).

A variety of statistical software packages are currently available for the analysis of multilevel data. Specialty packages, including HLM, MLWiN, and MIXREG, were designed for the treatment of multilevel data. Other packages have subroutines to handle multilevel analyses. These include LISREL, Mplus, SAS, SYSTAT, STATA, S-PLUS, and SPSS. We used the linear mixed procedure, PROC MIXED, available in SAS for the analysis of sentence processing data. PROC MIXED was developed for the analysis of a single continuous response variable assumed to follow a linear multilevel regression model. The procedure is flexible, allowing for a variety of error structures for repeated measures. Singer (1998) provides a tutorial for PROC MIXED for the treatment of clustered, multilevel data and growth models for longitudinal data. Readers are also referred to Littell, Milliken, Stroup, and Wolfinger (1996) for further details concerning the SAS mixed procedure.

Sentence Processing Experiment

We first consider data from a sentence processing experiment that tested readers' responses to sentences such as the following:

(1A) That's the daughter that the mother urged the doctor to cure yesterday.

(1B) That's the disease that the mother urged the doctor to cure yesterday.

Sentences such as (1A) and (1B) are temporarily ambiguous, because the initial noun phrase *daughter/disease* could serve as the direct-object argument of the main verb *urged*, but it actually serves as the object argument of the infinitival complement *to cure*. Verbs such as *urged* require two postverbal arguments, an object and an infinitival complement (e.g., *The mother urged the doctor*, by itself, is judged as being unacceptable; she has to urge the doctor to do something). Prior research suggests that read-

ers typically misanalyze sentences such as (1A) because they initially treat the first noun phrase (e.g., *the daughter*) as the object argument of *urged* (as in *The mother urged the daughter . . .*; Boland, Tanenhaus, Garnsey, & Carlson, 1995; Pickering & Traxler, 2001). When readers subsequently encounter the noun phrase *the doctor*, they experience a *filled gap* effect because they must undo the attachment of *daughter* to *urged* before they can form the correct association between *urged* and *doctor*. Comparable effects do not occur in sentences such as (1B), because the implausibility of *disease* as the direct object of *urged* and the availability of an obligatory infinitival complement causes readers to adopt the correct analysis immediately. Hence, based on previous eye-movement-monitoring experiments, sentences such as (1A) should be more difficult to process than sentences such as (1B), with the difference first appearing starting at the postverbal noun phrase *the doctor*.

The present experiment addresses the question of whether WM capacity, measured as a continuous variable, affects individual differences in the magnitude of disruption that readers experience in sentences such as (1A). An estimate of the degree of disruption can be obtained by comparing processing times for sentences such as (1A), where syntactic misanalysis occurs, with sentences such as (1B), where syntactic misanalysis does not occur or occurs to a much lesser extent. Given that recovery processes need to be undertaken for sentences such as (1A) but not for (1B), and assuming that parsing and reanalysis processes consume finite WM resources (as in the *shared resource* hypothesis; see, e.g., Just & Carpenter, 1992), readers with greater WM capacity should be less disrupted by the syntactic misanalysis in (1A) than readers with lesser WM capacity.

METHOD

Subjects

Sixty-one undergraduates from the University of California at Davis participated in return for course credit. All of the subjects had normal, uncorrected hearing and vision and were native speakers of English.

Stimuli

The test sentences were 28 pairs of items such as (1A) and (1B). The test sentences were rotated across lists such that equal numbers of each type appeared on each list and so that each subject saw exactly one version of each item. Presentation order was randomized individually for each subject. The test sentences were presented along with 52 items from two unrelated experiments and 30 filler sentences of various types.

Working Memory Assessment

The subjects completed a variant of the Daneman and Carpenter (1980) sentence span test (Turner & Engle, 1989). The subjects were presented with groups of sentences to read aloud. Each sentence was followed by an unrelated target word that the subjects were to remember.

After reading all of the sentences in a group, the subjects were asked to recall all of the target words in order. The number of sentences and the number of target words in a group increased from two to six as the subjects proceeded through the task. The subjects saw three groups of the same size before they saw the next larger group. They were initially given three groups of two sentences, as practice. We used raw scores (total number of words recalled in the correct serial position) as our estimate of WM capacity (see Waters & Caplan, 2003).

Eye-Movement-Monitoring Procedure

A Fourward Technologies dual-Purkinje-image eyetracker monitored the subjects' eye movements while they read the test sentences. The tracker has angular resolution of 10' of arc. The tracker monitored only the right eye's gaze location. A PC displayed materials on a VDU positioned 70 cm from the subject's eyes. The subject's gaze location was sampled every millisecond, and the PC software recorded the tracker's output to establish the sequence of eye fixations and their start and finish times. Before the experiment, the experimenter seated the subject at the eyetracker and used a bite plate and headrests to minimize head movements. After the tracker was aligned and calibrated, the experiment began. After reading each sentence, the subject pressed a key. After some of the filler sentences, the subject responded to a comprehension question. The subjects did not receive feedback on their responses. All of the subjects in the analyses reported below scored at 90% accuracy or above on the comprehension questions. Between each trial, a pattern of squares appeared on the computer screen along with a cursor that indicated the subject's current gaze location. If the tracker was out of alignment, the experimenter recalibrated it before proceeding with the next trial.

Scoring Region and Dependent Measure

Because our specific aim in this article is to contrast different analytic techniques, and in the interests of keeping the discussion simple, we report data from only one scoring region, consisting of the determiner and the noun that immediately followed the main verb (e.g., *the doctor*, in [1A] and [1B]). This region has previously been shown to effectively disambiguate the test sentences (Boland et al., 1995; Pickering & Traxler, 2001). Similarly, we only report *first-pass time*, which consists of all the fixations within the target region, beginning with the reader's first fixation and ending as soon as the reader's gaze leaves the region, either to the left or the right. Prior to analyzing the data, we eliminated any fixation times less than 120 msec or greater than 2,000 msec. Subsequently, we eliminated any remaining fixations that were greater than 2.5 *SDs* from the subjects' mean for that condition. These criteria resulted in 7.4% of the data being treated as missing.

A Traditional Approach Using ANOVA

The standard psycholinguistic approach to data such as those in Experiment 1 involves computing a mean response time (RT) for each subject for each condition and then analyzing those means using a within-subjects

ANOVA model. Extremely short (less than 120-msec) and extremely long (more than 1,000-msec) fixations are typically excluded prior to computing the means, as described above. When individual difference scores are obtained, they are typically used to group subjects, and group membership is then treated as a between-subjects (categorical) factor. For Experiment 1, the by-subject means were analyzed using a 2 (sentence type: sentences such as [1A] vs. sentences such as [1B]) $\times$ 2 (WM capacity: high vs. low, based on a median split, in this case) ANOVA model, with sentence type treated as a within-subjects factor and WM capacity treated as a between-subjects factor. Interest lies in the tests for the two main effects, sentence type and WM, and their interaction. Statistical tests were performed using a .05 level of significance.

Given one within-subjects factor and one between-subjects factor, each with two levels, the assumptions underlying valid inference from the results (described below) are as follows. For the test concerning the main effect of WM level, mean RTs aggregated across text levels are assumed independent and normal, with homogeneity of variance across WM levels. For the test concerning the main effect of the text factor, mean RTs across levels of WM are assumed normal within text levels, with homogeneity of variance assumed between text levels. For the interaction effect, mean RTs are assumed normal for each combined level of WM and text, with homogeneity of variance across combinations. Given that the present study involves a within-subjects factor with only two levels, issues concerning the assumption of sphericity do not apply. When there are more than two levels of a within-subjects factor, tests involving the within-subjects factor require that within each level of the between-subjects factor, the data satisfy the assumption of homogeneity of treatment-difference variances (see Maxwell & Delaney, 2004). Relevant assumptions were tested for the analyses (reported below) with satisfactory results.

The analyses resulted in an estimated mean RT of 439 msec for the noun phrase region in sentences such as (1A), and of 410 msec for the corresponding region of sentences such as (1B), collapsing across high- and low-span readers. This difference was statistically significant [$F(1,59) = 7.61$, $MS_e = 3,319$, $p = .008$]. The estimated mean first-pass time was 418 msec for high-span readers and 430 msec for low-span readers, collapsing across sentence types, a result that was not statistically significant [$F(1,59) = .449$, $MS_e = 10,517$, $p = .505$]. Likewise, the interaction between WM capacity and sentence type was not statistically significant [$F(1,59) = .038$, $MS_e = 3,319$, $p = .845$].

Subsidiary analyses conducted separately on each of the two WM groups showed that the sentence-type effect was significant for the low-span group [444 vs. 416 msec; $F(1,32) = 4.57$, $MS_e = 2,599$, $p = .040$], but not quite, for the high-span group [433 vs. 402 msec; $F(1,27) = 3.21$, $MS_e = 4,171$, $p = .084$]. Similar patterns have sometimes been interpreted as showing an effect of WM on sentence processing performance, although we would not like to make that claim here.

Multilevel Models and Analyses

In a multilevel model, the response is treated at the item level. Given repeated measures within subjects, data that vary from item to item are considered at the first level of the model, and data that vary across subjects are considered at the second level. In the first level, first-pass time is considered a function of sentence type. In addition to sentence type, we included a measure of region length to account for related RT differences within subjects. This information was not included in the ANOVA analyses presented above, because those analyses were based on subject-level means using aggregated scores.¹ Region length was centered about the sample mean length of 12.71 by taking the raw lengths and subtracting the mean length from each. This was helpful in the interpretation of the model below.

Using notation similar to that found in Raudenbush and Bryk (2002), the model at the first level is

$$\text{Level 1: } fp_{ij} = \beta_{0j} + \beta_{1j}(\text{SentenceType}_i) + \beta_{2j}(\text{RegionLength}_i) + \varepsilon_{ij},$$

where fp_{ij} is the first-pass RT on the i th item for subject j . The coefficient β_{0j} is the subject-level intercept for the regression of RT on sentence type (where sentence type is coded: 0 = sentences such as [1A] and 1 = sentences such as [1B]) and region length. Note that with this coding, the sentence-type coefficient is expected to be negative, because sentences such as (1B) should be *easier* to process than sentences such as (1A) in the region we analyzed. Due to the coding of sentence type and the centering of region length, the intercept of the Level 1 equation, β_{0j} , represents the expected RT for sentences such as (1A) for the j th individual for a region of average length. The coefficient β_{1j} represents the subject-level expected difference in RT for sentences such as (1B) versus sentences such as (1A), holding constant the effect of region length. The coefficient β_{2j} is the overall effect of the i th item's region length on RT. Under this model specification, the effect of region length is assumed constant across individuals. The error of the regression at this level of the model is denoted by ε_{ij} . The variance of this error term characterizes the within-subjects error variance after accounting for the predictors.

At the second level of the model, the coefficients at the first level are considered the criterion variables. Here, the subject-level intercept, β_{0j} , and sentence-type effect, β_{1j} , were allowed to vary as functions of WM. WM was centered about the sample mean by subtracting 39.13 from each subject's observed WM score. This was done to improve interpretation of the intercepts of each Level 2 regression, as described below. The regressions of the Level 1 coefficients at the second level are specified as

$$\begin{aligned} \text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}\text{WM}_j + r_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\text{WM}_j + r_{1j} \\ \beta_{2j} &= \gamma_{20}. \end{aligned}$$

In these regressions, the coefficients γ_{00} and γ_{10} represent the overall RT for sentences such as (1A) and the overall difference in RT for sentences such as (1B) versus (1A),

respectively, when WM is at the sample mean. In the regression of β_{0j} , the coefficient γ_{01} represents the effect of WM scores on a subject's average RT for sentences such as (1A). Error in the subject-level mean RT for sentences such as (1A) not accounted for by WM is captured by the term r_{0j} . In the regression of β_{1j} , the coefficient γ_{11} represents the moderating effect of WM on the regression of RT on sentence type. That is, γ_{11} is the cross-level interaction effect between WM and sentence type on RT. Error in the subject-level effect of sentence type on RT not accounted for by WM scores is represented by the term r_{1j} . The final equation is for the effect of region length on processing time, β_2 . This effect is assumed constant across subjects, so it is expressed as a direct function of the constant γ_{20} .

Parameter Interpretation

In estimating the model, the error at the first level and the subject-specific coefficients at the second level are not considered directly. Rather, the model generates a mean and covariance structure for the responses (see Raudenbush & Bryk, 2002). The model's mean structure represents the response at the population level, and involves the fixed parameters γ_{00} , γ_{01} , γ_{10} , γ_{11} , and γ_{20} , as described above. The model's covariance structure represents the expected value of the variance-covariance matrix for the observed responses, based on assumptions about the errors at both levels of the model. In a multilevel model, the covariance matrix is decomposed into a within-subjects part and a between-subjects part. The within-subjects component is characterized by the variance of the within-subjects error γ_{ij} . In many cases, the within-subjects errors are assumed to be independent and identically distributed as normal, with a mean of zero and variance σ^2 . For the present study, this would mean that the error variance is assumed to be constant across items and individuals.

Corresponding to the second level of the model, the variances of the random coefficients (the intercept and the slope relating to the effect of region length on RT) and their covariance help characterize the second component of the decomposed variance-covariance matrix based on model assumptions. Here, the two random coefficients, r_{0j} and r_{1j} , each have a variance, ϕ_0 and ϕ_1 , respectively, and together have a covariance, ϕ_{10} . The variances represent individual differences in the intercept and slope after taking into account WM scores, as described above in the Level 2 regressions of the Level 1 intercept and slope. Assuming constant error variance at the first level across items and subjects, the fixed effects in addition to the variances and covariance relating to the random parts of the

model yield a total of nine parameters: γ_{00} , γ_{01} , γ_{10} , γ_{11} , γ_{20} , ϕ_{00} , ϕ_{10} , ϕ_{11} and ϕ^2 . We used SAS Version 9.1 to estimate the model parameters based on the sample data.

Variance and Covariance Parameter Estimates for the Second Level Model

We consider the estimates for the variances and covariance corresponding to the random coefficients at the second level of the model (see Table 1). Relating to each parameter is an estimate and its corresponding asymptotic standard error. The ratio of an estimate to its standard error is a Wald Z statistic. Its significance value, corresponding to a null hypothesis that the statistic is equal to zero, may be based on a one-tailed test, for tests concerning variance estimates, or a two-tailed test, for tests concerning covariance estimates (see Self & Liang, 1987). Inference is based on the assumption that the parameter estimate is normally distributed. The sampling distribution of a sample variance or covariance is considered normal for large sample sizes and when the true value is not close to zero. In the present study, the sample size was $N = 61$; consequently, inference based on a Wald Z statistic is not likely to be reliable.

A preferred procedure for evaluating the necessity of the random coefficients at the second level is a likelihood ratio test. A likelihood ratio statistic is obtained by taking the difference between the deviances (i.e., $-2 * \log$ likelihood) corresponding to each model. The difference is distributed approximately as chi-square, for large samples, with degrees of freedom equal to the difference in the number of parameters used to characterize the two models. The lack of a statistically significant difference would essentially suggest that the simpler model provides about as good a fit as the more complicated model, but does so with fewer parameters. As a method for evaluating the necessity of the random coefficients at the second level, this approach is considered conservative, in that the probability of rejecting a false null hypothesis is reduced (Pinheiro & Bates, 2000).

As a means of comparison with the model that included a random intercept and slope and their covariance, we fitted a restricted model in which the random slope was excluded, resulting in the estimation of two fewer parameters (i.e., the variance of the random slope and the covariance between the random constant and slope were set equal to zero). The deviance for the least-restrictive model was 20,547.9. The resulting deviance for the restricted model was 20,555.7. With a difference of 2 *df* between the models, the difference in deviances of 7.8 was statistically

Table 1
Variances and Covariance for Random Coefficients at Level 2

Parameter	REML	SE	Z	p
Random constant variance, ϕ_0	5,663	1,459	3.88	<.0001
Random slope variance, ϕ_1	2,233	1,224	1.82	.034
Random constant and slope covariance, ϕ_{10}	-2,273	1,116	-2.04	.042

Note—*p* is based on a one-tailed test for variance estimates and a two-tailed test for the covariance estimate. REML, restricted maximum likelihood estimate; Z, the ratio of an estimate to its asymptotic standard error, valid only for large samples.

significant ($p = .020$), suggesting that the random slope and covariance between it and the intercept resulted in an improvement in the overall model fit. Assuming reliable random variation in the slope associated with the effect of sentence type on RT, these results suggest that individual differences in the item effect on RT remain, after taking WM into account.

Corresponding to the covariance between the intercept and slope was a correlation of $r = -.64$, suggesting a moderate negative association between RT for sentences such as (1A) and the effect of sentence type on RT when WM was at the sample mean. This suggests that subjects whose RTs are relatively high for sentences such as (1A) tend to show stronger, negative effects of sentence type on RT. Interpretation of the Level 1 intercept is dependent on how the predictors are coded. Here, sentence type was coded as 0 = sentences such as (1A), 1 = sentences such as (1B), and WM and region length were centered to their sample means. This coding had consequences on the interpretation of the correlation between the intercept and the effect of sentence type on RT. That is, if the coding of sentence type had been reversed (as 0 = sentences such as [1B]; 1 = sentences such as [1A]), the correlation between the intercept and the slope would become the correlation between RT for sentences such as (1B) and the effect of sentence type on RT for individuals when WM was at the sample mean. After recoding sentence type and reestimating the model, the estimated correlation between the intercept and slope was $r = .01$, a value that was small and statistically not different from zero, suggesting no association between RT for sentences such as (1B) and the effect of sentence type on RT. Similarly, the WM measure and region length could be centered to other values, such as those that might be of theoretical interest.

Fixed-Effects Estimates

Estimates for the fixed effects with their standard errors and the corresponding test statistics are presented in Table 2. The estimated overall RT for sentences such as (1A) was 439.22 when WM and length were each at their sample means of 39.13 and 12.71, respectively. The test statistic associated with this estimate is based on a null hypothesis that the parameter is equal to zero, which in the present context does not provide a useful result.

We then evaluate the effect of sentence type on RT for sentences such as (1B) versus sentences such as (1A), holding constant the effects of WM and region length. The estimated effect of -29.8 was large relative to its standard

error of 10.5, suggesting that the true sentence-type effect was different from zero and that RT overall was reduced for items such as (1B). Further, a 95% confidence interval for the difference in RT between items such as (1A) and (1B) was $(-50.4, -9.22)$, indicating a wide range of plausible mean differences. Note that this sentence-type effect (with sentences such as [1B] being easier to process than sentences such as [1A] in the region we analyzed) replicates findings from previous studies (Boland et al., 1995; Pickering & Traxler, 2001).

The point estimate for the effect of WM on RT when subjects were responding to sentences such as (1A) was -2.14 with a standard error of 1.67. The corresponding 95% confidence interval estimate was $(-5.41, 1.13)$, an interval that included zero as an interior value. The estimated difference in the effect of WM between sentences such as (1A) and (1B) was .632, with a relatively large standard error of 1.53. A 95% confidence interval of $(-2.37, 3.63)$ included zero as an interior value, suggesting no difference between sentences such as (1A) and (1B) in the effect of WM on RT. Finally, the effect of region length was 8.90 with a standard error of 2.11, suggesting a positive association with processing time. A 95% confidence interval for the effect was $(4.76, 13.0)$.

DISCUSSION

Studies looking for individual differences in sentence processing performance have relied heavily on ANOVA modeling techniques. Valued for its simplicity, ANOVA is effective in detecting mean differences among groups that are defined by levels of a nominal-level variable. A common approach in this research domain is to create groups by categorizing a continuous measure, such as WM, and then to compare the groups with regard to mean processing times. Problems with this general strategy of creating categories from continuous predictors have been well documented in the methodological literature (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; also see Preacher, Rucker, MacCallum, & Nicewander, 2005, for a discussion concerning extreme-group analyses), although the practice continues.

The use of multilevel models for repeated measures is increasingly common in the social and behavioral sciences. These models handle both continuous and categorical predictor variables and are most commonly applied to continuously distributed response variables. Relative to classic approaches such as ANOVA, multilevel models are flexible, allowing tests of effects not handled by classic methods. In the context of a sentence processing experiment, a multilevel model allows for predictors to be measured at both the sentence and subject levels, and the interactions between variables at different levels may also be examined. In the present application, we considered region length at the item level, as well as the interaction between sentence type (a sentence-level variable) and WM (a subject-level variable).

Using a multilevel model, we could allow for individual differences in the effect of sentence type on RT. This could not be tested in the standard ANOVA approach, which

Table 2
Estimated Fixed Effects for the Multilevel Model

Parameter	REML	SE	<i>t</i>	<i>p</i>
Intercept, γ_{00}	429	10.4	41.4	<.0001
WM, γ_{01}	-.958	1.51	-0.64	.527
Sentence type, γ_{10}	-29.1	9.77	-2.98	.003
WM $\times$ sentence type, γ_{11}	-.547	1.42	-.038	.701
Region length, γ_{20}	9.91	1.96	5.07	<.0001

Note—The *t* value is the ratio of the estimate to its standard error; *p* values are based on two-tailed tests. REML, restricted maximum likelihood estimate.

would assume that the effect is constant across individuals. When the effect is random, there may overall be a positive, negative, or no effect of sentence type on processing time. With the inclusion of a random coefficient for the item characteristic, the item effect can deviate about the average. If the overall effect of the sentence-type manipulation is positive, for example, then significant variation in the effect at the subject level would imply that for some subjects, the effect of the sentence-type manipulation is stronger than the average; for others, the effect is weaker. In another example, if the overall effect is null, then significant variation in the effect at the subject level would imply that for about half of the subjects, the effect of the item characteristic is positive, and for about the other half, the effect is negative.

Given subject-to-subject variation in item characteristic effects, tests of the moderating effects of subject-level variables (e.g., WM capacity) on the subject-level effects measured at the first level can be tested by way of regressions at the second level of the model. Subject-level variables that are included in the second level regression of the Level 1 intercept allow for tests of the effects of such variables on the mean subject-level response when all other variables in the first level of the model are equal to zero. Subject-level variables that are included in the second level regression of the Level 1 slope associated with a Level 1 predictor allow for tests of the moderating effects of subject-level variables on the effect of that Level 1 predictor. These moderating effects are also referred to as cross-level interaction effects, and represent one benefit of the multilevel approach.

This analytic strategy also easily handles incomplete data without imputation. That is, analyses are based on whatever data are available for each subject, with no requirement that subjects have the same number of observations. In cases of incomplete data, inference from the estimated model is dependent on the status of the *missingness*. That is, the model as presented here is based on the assumption that data are missing at random, such that the missingness is independent of the missing data. A simple method to assess the assumption of data missing at random is to apply a pattern-mixture random-effects model to the data (see, e.g., Hedeker & Gibbons, 1997). In this approach, one creates indicator variables to represent the different patterns of missing data in the response variable. These indicator variables are then included in the model as predictors. In this way, one can test the effect of the missing data. When the pattern indicators are judged reliable, and assuming that the appropriate patterns have been represented, they are retained in the model, and the assumption of missing at random may be satisfied.

There are many different commercially available software packages for multilevel analyses. Specialized packages include HLM, MLwin, and Hedeker's programs, which can be downloaded free of charge at his Web site. Widely used commercial software packages that can also handle a variety of multilevel models include LISREL, Mplus, SYSTAT, SPLUS, SAS, and SPSS. These packages vary in terms of the kinds of data they can handle (i.e., continuous vs. categorical outcome data), the kinds

of models (e.g., linear vs. nonlinear) that can be fitted, and flexibility with regard to the kinds of model assumptions made (e.g., independent subject-level error with constant variance vs. a first-order auto-regression error structure). We used the SAS procedure, PROC MIXED, because it is generally widely available and commonly used by researchers in the social and behavioral sciences, and was well suited to the analysis of data from the present study.

AUTHOR NOTE

This research was supported in part by Grants R01-40865 from the National Institute of Child Health and Human Development, 0446618 from the National Science Foundation, and R01-HD048914-01A2 from the National Institutes of Health (all awarded to the second author). The authors thank Jason Golubock and Lori Miyasato for assistance in collecting the data. Correspondence concerning this article should be addressed to S. A. Blozis, Psychology Department, University of California, One Shields Avenue, Davis, CA 95616 (e-mail: sablozis@ucdavis.edu).

REFERENCES

- BADDELEY, A. [D.] (1986). *Working memory*. Oxford: Oxford University Press.
- BADDELEY, A. D., & HITCH, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47-90). San Diego: Academic Press.
- BOLAND, J. E., TANENHAUS, M. K., GARNSEY, S. M., & CARLSON, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh- questions. *Journal of Memory & Language*, **34**, 774-806.
- CAPLAN, D., & WATERS, G. S. (1990). Short-term memory and language comprehension: A critical review of the neuropsychological literature. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 337-389). Cambridge: Cambridge University Press.
- CAPLAN, D., & WATERS, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral & Brain Sciences*, **22**, 77-126.
- CLIFTON, C., JR., TRAXLER, M. J., MOHAMED, M. T., WILLIAMS, R. S., MORRIS, R. K., & RAYNER, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory & Language*, **49**, 317-334.
- COHEN, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, **7**, 249-253.
- DANEMAN, M., & CARPENTER, P. A. (1980). Individual differences in comprehending and producing words in context. *Journal of Verbal Learning & Verbal Behavior*, **19**, 450-466.
- HEDEKER, D., & GIBBONS, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, **2**, 64-78.
- JUST, M. A., & CARPENTER, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory capacity. *Psychological Review*, **99**, 122-149.
- JUST, M. A., CARPENTER, P. A., & KELLER, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, **103**, 773-780.
- KING, J. W., & JUST, M. A. (1991). Individual differences in syntactic parsing: The role of working memory. *Journal of Memory & Language*, **30**, 580-602.
- LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., & WOLFINGER, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.
- LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Association*, **90**, 1112-1121.
- 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. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- PEARLMUTTER, N. J., & MACDONALD, M. C. (1995). Individual differ-

- ences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory & Language*, **34**, 521-542.
- PICKERING, M. J., & TRAXLER, M. J. (2001). Strategies for processing unbounded dependencies: Lexical information and verb-argument assignment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1401-1410.
- PINHEIRO, J. C., & BATES, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- PREACHER, K. J., RUCKER, D. D., MACCALLUM, R. C., & NICEWANDER, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, **10**, 178-192.
- RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- SELF, S. G., & LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605-610.
- 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.
- SNIJDERS, T. A. B., & BOSKER, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- TRAXLER, M. J., MORRIS, R. K., & SEELY, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory & Language*, **47**, 69-90.
- TRAXLER, M. J., PICKERING, M. J., & CLIFTON, C., JR. (1998). Syntactic parsing is not a form of lexical ambiguity resolution. *Journal of Memory & Language*, **39**, 559-582.
- TRAXLER, M. J., WILLIAMS, R. S., BLOZIS, S. A., & MORRIS, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory & Language*, **53**, 204-224.
- TURNER, M. L., & ENGLE, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, **28**, 127-154.
- VERSCHUEREN, N., SCHAEKEN, W., & D'YDEWALLE, G. (2005). Everyday conditional reasoning: A working memory-dependent tradeoff between counterexample and likelihood use. *Memory & Cognition*, **33**, 107-119.
- WANNER, E., & MARATSOS, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119-161). Cambridge, MA: MIT Press.
- WATERS, G. S., & CAPLAN, D. (1996). Processing resource capacity and the comprehension of garden path sentences. *Memory & Cognition*, **24**, 342-355.
- WATERS, G. S., & CAPLAN, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, **35**, 550-564.

NOTE

1. One approach to taking the effect of sentence length on RT into account in an ANOVA is to regress RT on region length and then aggregate the resulting residuals to form the data upon which the analysis will be done. In eye-movement studies, this procedure is typically called a *residual reading time analysis* (see, e.g., Traxler, Pickering, & Clifton, 1998).

(Manuscript received May 10, 2005;
revision accepted for publication November 11, 2005.)