Tutorial

How Mixed-Effects Modeling Can Advance Our Understanding of Learning and Memory and Improve Clinical and Educational Practice

Katherine R. Gordon^a

Purpose: A key goal of researchers, clinicians, and educators within the fields of speech, language, and hearing sciences is to support the learning and memory of others. To do so, they consider factors relevant to the individual, the material to be learned, and the training strategy that can maximize learning and retention. Statistical methods typically used within these fields are inadequate for identifying the complex relationships between these factors and are ill equipped to account for variability across individuals when identifying these relationships. Specifically, traditional statistical methods are often inadequate for answering questions about special populations because samples drawn from these populations are usually small, highly variable, and skewed in distribution. Mixed-effects modeling provides advantages over traditional statistical techniques to answer complex questions while taking into account these common characteristics of special populations.

Method and Results: Through two examples, I illustrate advantages of mixed-effects modeling in answering

questions about learning and memory and in supporting better translation of research to practice. I also demonstrate key similarities and differences between analysis of variance, regression analyses, and mixed-effects modeling. Finally, I explain three additional advantages of using mixed-effects modeling to understand the processes of learning and memory: the means to account for missing data, assess the contribution of variations in delay intervals, and model nonlinear relationships between factors.

Conclusions: Through mixed-effects modeling, researchers can disseminate accurate information about learning and memory to clinicians and educators. In turn, through enhanced statistical literacy, clinicians and educators can apply research findings to practice with confidence. Overall, mixed-effects modeling is a powerful tool to improve the outcomes of the individuals that researchers and practitioners serve within the fields of speech, language, and hearing sciences.

Clinicians and educators understand the importance of maximizing the efficiency and effectiveness of learning in the individuals they serve. For clinicians, the increased number of individuals who qualify for speech and language services in addition to a shortage of speech-language pathologists has led to increased caseloads (Edgar & Rosa-Lugo, 2007; Katz, Maag, Fallon, Blenkarn, &

Smith, 2010). Thus, clinicians are invested in implementing treatments that will lead to the best long-term outcomes given the limited time they have with each client. Educators are expected to teach a large amount of information and skills across a variety of subject areas to prepare students for success in postsecondary education and workplace settings (American Speech-Language-Hearing Association, 2010; Common Core State Standards Initiative, 2010; Every Student Succeeds Act, 2015; Powell, 2018). Given this immense challenge, they are invested in teaching in an efficient way that will lead to both effective learning and retention of taught material.

Researchers within the fields of speech, language, and hearing sciences are invested in identifying how to improve learning and memory in the populations that clinicians and educators serve. However, significant barriers remain in translating clinical and educational research to real-world

^aCenter for Childhood Deafness, Language, and Learning Research, Boys Town National Research Hospital, Omaha, NE

Correspondence to Katherine R. Gordon:

Katherine.Gordon@boystown.org

Editor-in-Chief: Ryan McCreery

Editor: Frederick Gallun

Received June 15, 2018

Revision received November 5, 2018

Accepted December 10, 2018

https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0240

Publisher Note: This article is part of the Research Forum: Advancing Statistical Methods in Speech, Language, and Hearing Sciences.

Disclosure: The author has declared that no competing interests existed at the time of publication.

settings. Identifying the most effective and efficient way to teach can be complex as characteristics of the individual (McGregor, Gordon, Eden, Arbisi-Kelm, & Oleson, 2017; Storkel, Komesidou, Fleming, & Romine, 2017), the material to be learned (Storkel, 2018), and the training strategy (Nunes & Karpicke, 2015; Rowland, 2014) can all affect the likelihood that material will be learned and retained. Relevant learner factors include age, current communication abilities, current cognitive abilities, and any specific disorders or disabilities. Characteristics of the material to be learned include the subject area (e.g., science, math), the medium through which the information is conveyed (e.g., oral, written), and the type of information (e.g., facts or procedures). Characteristics of the training strategy include the type of learning activity. For example, the learner could be listening to a science lesson while taking notes, or they could be participating in a group activity facilitated by the teacher. Characteristics of the training strategy also include the intensity with which learning material is targeted, namely, how many times a week the learner focuses on that information, and the dosage of the target material during each learning session, such as how many times they hear a specific vocabulary word (Alt, Meyers, & Ancharski, 2012; Justice, Logan, Schmitt, & Jiang, 2016).

The researcher's role in addressing these complex questions is to discover and communicate accurate information about learning and retention. By using effective statistical tools, researchers can improve the quality of the information they give to practitioners, which, in turn, will improve their ability to translate research findings to real-world practice. To be effective, these tools must allow researchers to identify how each relevant factor contributes to the success of learning and retention while taking into account individual differences. Given that learning is inherently a longitudinal process, these tools must allow researchers to determine how memory for information changes throughout learning and over delays. As most longitudinal research involves attrition, these tools must allow researchers to maximize the usefulness of the data that they have while taking into account missing data. Finally, given that learning often does not occur at a consistent rate over time (Law, Tomblin, & Zhang, 2008; Rice & Hoffman, 2015), these tools must allow researchers to model nonlinear relationships between various factors and learner outcomes.

Mixed-effects modeling is one such tool. Mixed-effects modeling, also known as multilevel modeling, hierarchical linear modeling, and random coefficient modeling, has gained favor in the social sciences in recent years (Baayen, Davidson, & Bates, 2008; Gries, 2015; Locker, Hoffman, & Bovaird, 2007; Quené & van den Bergh, 2004) and offers some unique benefits to the fields of speech, language, and hearing sciences. Speech-language pathologists, audiologists, and many educators work with people from special populations that are diverse (e.g., autism spectrum disorder; Waterhouse, 2013) and not normally distributed (e.g., intellectual disabilities; Maulik, Mascarenhas, Mathers, Dua, & Saxena, 2011). However, commonly used statistical techniques, such as analysis of variance (ANOVA), are often

ill equipped to deal with special populations that include a high degree of variability across participants and nonnormal distributions (Baayen et al., 2008; Gueorguieva & Krystal, 2004). Critically, mixed-effects modeling allows researchers not only to control for individual variability when identifying relationships between predictors and outcomes but also to identify how the variability in individual characteristics, learning material, and training strategies specifically contributes to learner outcomes (Misangyi, LePine, Algina, & Goeddeke, 2006). This is vital to improving clinical and educational practice given the variability in individual responses to interventions and classroom instruction (Kelley, Leary, & Goldstein, 2018).

Mixed-effects modeling provides benefits for answering questions across a variety of topic areas within the fields of speech, language, and hearing sciences (see Perry & Kucker, 2019; Walker, Redfern, & Oleson, 2019, in this issue). However, the purpose of this tutorial is to illustrate how mixed-effects modeling is an effective statistical tool to answer questions about learning and memory, processes important for achieving a wide variety of clinical and educational goals. I provide two examples. In the first example, I illustrate the utility of ANOVA, regression analyses, and mixed-effects modeling to determine how individual factors contribute to learner outcomes. In the second example, I illustrate how mixed-effects modeling is an effective tool to identify efficient training strategies that maximize retention. For each example, I discuss clinical and educational implications of using mixed-effects modeling to address these questions. Additionally, I explain three unique benefits of mixed-effects modeling, which include the means to account for missing data, assess the contribution of variations in delay intervals, and model nonlinear relationships between factors.

Example 1: How Individual Factors Affect Learner Outcomes

Let us consider an example data set from my own research (Gordon et al., 2016) and compare the usefulness of different statistical approaches in determining how individual factors influence learner outcomes. In Gordon et al. (2016), 3-year-old and 4- to 5-year-old children with typical development participated in four sessions across two weeks (see Table 1). During the first week, they were trained on a set of six word-object pairs (e.g., Set A) and were tested on those pairs at the end of training (i.e., short-term test) and two days later (i.e., long-term test). During the second week, they were trained on an additional set of six word-object pairs (e.g., Set B) and tested on that set at the end

Table 1. Training and testing schedule for each participant in Gordon et al. (2016).

Week	End of training	Two days after training
Week 1	Short-term test (e.g., Set A)	Long-term test (e.g., Set A)
Week 2	Short-term test (e.g., Set B)	Long-term test (e.g., Set B)

of training and two days later. Order of object sets (A, B) was counterbalanced across subjects. To control for children's experience with both the words and their referents, we used objects that were unfamiliar to children and created novel labels for each object (e.g., dorb, gramer).

During both the short- and long-term tests, children were shown trained objects one at a time and asked, "What is this one called?". They were given three word forms to choose from: the target form (e.g., dorb), a minimal pair of the target (e.g., vorb), and a maximally different form (e.g., zinnip). In the following analyses, I focus on whether children correctly identified the target form or not. Object set did not significantly predict children's responses, so this factor was eliminated from the analyses. Thus, the predictor variables of interest were delay (short-term, long-term), age (younger, older), receptive vocabulary (raw score from the Peabody Picture Vocabulary Test [PPVT]–Fourth Edition [Dunn & Dunn, 2007]), and week (1, 2), and the outcome variable was whether children selected the correct forms for the objects.

In the following analyses, I assessed the degree to which each predictor variable contributed to children's responses in the word learning tests. If any of these variables had a significant effect on the outcome variable, it would be considered a main effect (Hox, Moerbeek, & van de Schoot, 2017). For example, a main effect for vocabulary might be that children with a higher PPVT score identified more objects correctly than children with a lower PPVT score. I also assessed whether there were any significant interactions between variables. An interaction occurs when the relationship between one predictor variable (e.g., week) and the outcome differs based on another predictor variable (e.g., age; Hox et al., 2017). An example of an interaction would be an Age $\times$ Week interaction if younger children perform better during Week 1 than Week 2, but older children perform similarly across weeks.

ANOVA

As is common for an ANOVA, I coded performance at the short- and long-term tests as the total number of items correct. I then analyzed the data via a Delay (short-term, long-term) $\times$ Week (1, 2) $\times$ Age (younger, older) $\times$ Vocabulary (low, high) mixed repeated-measures ANOVA. The ANOVA is mixed because it includes both between-subject and within-subject factors. Note that the term *mixed* means something quite different when referring to mixed-effects modeling, which will be described below. Age and vocabulary group were between-subject factors in that each participant only had one age and were only part of one vocabulary group. To assign vocabulary group, I used a median split of raw PPVT scores within each age group. Week and delay were within-subject factors in that each participant participated in testing across multiple weeks and multiple delay intervals. I used a repeated-measures ANOVA, as opposed to a traditional ANOVA, because each child was tested multiple times in the word learning task.

Results from the repeated-measures ANOVA are shown in Table 2. This analysis revealed a significant main

Table 2. Results from the repeated-measures analysis of variance.

Predictor	$F(1, 28)$	p	η_p^2
Vocab	0.01	.10	.09
Age	4.02	.06	.13
Vocab $\times$ Age	2.84	.10	.09
Delay	0.57	.46	.02
Delay $\times$ Vocab	5.17	.03*	.16
Delay $\times$ Age	0.14	.71	.01
Delay $\times$ Vocab $\times$ Age	0.57	.46	.02
Week	31.18	< .001**	.53
Week $\times$ Vocab	0.02	.90	.01
Week $\times$ Age	0.83	.37	.03
Week $\times$ Delay	0.00	1.00	.00
Week $\times$ Delay $\times$ Vocab	0.14	.71	.01
Week $\times$ Delay $\times$ Age	1.23	.28	.04
Week $\times$ Vocab $\times$ Age	0.42	.52	.02
Week $\times$ Delay $\times$ Vocab $\times$ Age	0.14	.71	.01

Note. Age and vocabulary (vocab) group were within-subject factors. Delay and week were between-subject factors.

* $p < .05$. ** $p < .001$.

effect for week, $F(1, 28) = 31.18$, $p < .001$, with children performing better during Week 1 than Week 2. The main effect for age approached but did not reach significance, $F(1, 28) = 4.02$, $p = .06$, with older children performing better than younger children. There was a significant Delay $\times$ Vocabulary Group interaction, $F(1, 28) = 5.17$, $p = .03$. To follow up on this interaction, I conducted a separate repeated-measures ANOVA for each vocabulary group with delay (short-term, long-term) as the within-subject factor and data combined across weeks (see Figure 1). These analyses revealed that children in the low-vocabulary group performed worse at the long-term tests than the short-term tests, $F(1, 14) = 4.90$, $p = .04$. There was not a significant effect of delay for the high-vocabulary group, $F(1, 14) = 1.08$, $p = .32$.

Figure 1. Mean number of words children correctly identified at the short- and long-term tests in the low- and high-vocabulary groups. The total is out of 12 as responses were collapsed across weeks.

What can I conclude from these analyses about how individual factors affect learner outcomes? The primary conclusion is that children with a lower vocabulary forgot some of the information they learned after a two-day delay. In contrast, children with a higher vocabulary, by and large, retained the information they learned. These conclusions are sensible. However, how much information are children with a lower vocabulary likely to forget? It is hard to say exactly from these results. Children in the low-vocabulary group forgot, on average, 1.00 word ($SD = 1.79$) over the delay. However, that average disguises a range of performance from one child who forgot four items to another child who correctly identified two additional items after the delay. In contrast, the high-vocabulary group performed slightly better at the long- than short-term test, an average nonsignificant improvement of 0.5 word ($SD = 1.86$) with a range of -4 to $+3$ words. Given this information, could a clinician predict how many words a child with a given age and receptive vocabulary would remember two days after training? Based on the findings from the current analyses, they would have a general idea, but it would be difficult to make a specific prediction.

Mixed-Effects Modeling

Mixed-effects models are similar to regression analyses in that they are used to determine an equation (i.e., a model) that best fits the relationship between the predictor variables (e.g., age, vocabulary) and the outcome variable (e.g., performance in the word learning tests; Gelman & Hill, 2006). This relationship can be either linear or nonlinear (Mirman, 2016). Linear relationships are found when a change in the predictor variables leads to a systematic change in the outcome variable. Nonlinear relationships are found when the effect that the predictor variables have on the outcome variable is inconsistent, such as a child showing a slow language growth rate at the beginning of an intervention but showing an increased growth rate after some time in the intervention. I will give more details about nonlinear relationships and models used for these analyses in a later section. For a linear relationship, the model has the following structure:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + e \quad (1)$$

For my data, the model will have the following structure if I include all possible main effects.

$$Y = \beta_0 + (\beta_1 \times \text{Delay}) + (\beta_2 \times \text{Week}) + (\beta_3 \times \text{Age}) + (\beta_4 \times \text{PPVT}) + e \quad (2)$$

Y represents performance in the word learning test for a given participant based on the delay (e.g., short-term, long-term), week of assessment, age group, and PPVT score. For this analysis, I coded delay, week, and age dichotomously: delay (short-term = 0, long-term = 1), week (Week 1 = 0, Week 2 = 1), and age (younger = 0, older = 1). However, I retained receptive vocabulary score as a continuous variable and centered the mean raw PPVT score at 0

in the model. Hox et al. (2017) provide an explanation of the advantages of centering data at the mean. I could have maintained age as a continuous factor in the model, which cannot be done in ANOVA. It is useful to do so if the researcher is interested in how changes in age on a continuous scale affect change in an outcome variable. For example, early in development children often go through rapid changes over very short time scales. Thus, including age in months or even age in days as a continuous variable might add useful insights into the effect of age on a specific outcome. Additionally, if we were considering a broad age range such as individuals from 20 to 80 years old, including age as a continuous variable could provide valuable information about how age is related to change in an outcome variable. For my current data set, I am interested in how one age group differs from another age group, so I maintained age as a dichotomous variable.

In the model, β_0 represents the intercept—the predicted outcome when all other predictors are 0. In other words, β_0 represents the predicted performance in the word learning task at the short-term delay (short-term = 0) during the first week (Week 1 = 0) of a child from the younger age group (younger = 0) with a mean PPVT score (the mean is centered at 0). β_1 , β_2 , β_3 , and β_4 represent coefficients for the delay, week, age, and PPVT score predictors, respectively (Gelman & Hill, 2006). The coefficients indicate how much the outcome variable changed relative to a change in that predictor variable. For example, the coefficient for week would reflect how much a change from the first week to the second week affected children's performance in the word learning tests. Any regression line we fit to the data is not going to fit all data points exactly. Thus, we need to include an error term in the model (e) to account for the distance between the regression line and each individual data point (i.e., the residual). In other words, e represents the difference between how we would predict a given child would perform based on the model and how they actually performed (see Fielding & Goldstein, 2006).

Changing the Outcome Variable to the Probability of a Correct Response

Considering my data, it is important to think about the outcome variable I want to use in my model. It could be the number of items that a child correctly identified at the short- and long-term tests as it was for the ANOVA. However, by summing the data in this way, I am missing some key information. Whether a child responded correctly to a specific item (e.g., the object labeled a “dorb”) at the short-term test is likely a good predictor of whether they responded correctly to that item at the long-term test. On the surface, predicting the probability of a correct response to a specific item as opposed to the total number of items correct may not seem that different. However, suppose a clinician was training children to remember and produce six new vocabulary words using pictures. They notice that two different children consistently named three pictures correctly during the last four training sessions. On closer

inspection, they notice that one child responded correctly to the same three pictures each time, but the other child varied in which pictures they named correctly. These two patterns of responses suggest different learner outcomes. The first child learned three words very well but has yet to learn the other three words. The second child has some memory for all of the words, but these memories are not strong enough to support consistent responses. Given this information, the clinician would likely use different strategies for each child to support the learning of all six words. To fully understand and support learning and memory processes, it is often important to analyze data on the item level.

To assess the effects of individual factors (e.g., age, vocabulary) on memory for items, children's responses to specific items (e.g., the object labeled a "dorb") at the short-term test are included in the model to predict their response to that item at the long-term test. Now, the predicted outcome (Y) is no longer the number of items correct but the probability (p) that a child will respond correctly to a specific item at the long-term test. This variable will range from 0 to 1—in other words, 0%–100% probability. After I make these changes, the model has the following structure:

$$p = \beta_0 + (\beta_1 \times \text{Short-Term Response}) + (\beta_2 \times \text{Week}) + (\beta_3 \times \text{Age}) + (\beta_4 \times \text{PPVT}) + e \quad (3)$$

Although it is an advantage to predict the probability of a correct response to each item, some problems remain with this model (see Jaeger, 2008). First, because the outcome is bounded by 0 and 1, there is likely more variation around values close to 0.5 and less variation around values close to 0 and 1. This is because the probability of a correct response can vary both above and below 50%. However, the probability of a correct response can only occur above 0% and below 100%. For example, the probability that four participants identified the correct label for the object called a dorb might be .85, .90, .95, and .97. The probability will never be 1.10 or 110%. This violates the assumption of homogeneity of variance inherent to these analyses (Jaeger, 2008). I can address this problem by transforming the outcome variable to the odds of a correct response through the following formula:

$$p/(1 - p) \quad (4)$$

In this case, the probability of a correct response is divided by the probability of an incorrect response. Now, the outcome variable ranges from 0 to positive infinity, and an odds of 1 represents a 50% probability of a correct response (i.e., equal odds) as follows:

$$.50/(1 - .50) = 1 \quad (5)$$

Although I addressed the problem of heterogeneity of variance, there is a second problem. Odds are not linearly related to predictor variables. By taking the natural logarithm

of odds, which is called the logit or the log odds, the relationship between my predictor variables and outcome variable becomes much more linear (Jaeger, 2008). Thus, my model now has the following structure:

$$\ln(p/1 - p) = \beta_0 + (\beta_1 \times \text{Short-Term Response}) + (\beta_2 \times \text{Week}) + (\beta_3 \times \text{Age}) + (\beta_4 \times \text{PPVT}) + e \quad (6)$$

Now, my data range from negative to positive infinity and are centered around 0. It should be noted that these types of changes are important to avoid spurious results when the outcome variable is dichotomous (e.g., correct = 1, incorrect = 0) as opposed to continuous (e.g., a range from zero to six items correct), which is common in the social sciences. Johnson (2014), Morrison and Kondaurova (2009), Quené and van den Bergh (2008), and Storkel, Bontempo, Aschenbrenner, Maekawa, and Lee (2013) provide examples of data with dichotomous outcome variables analyzed with a logit transformation.

Adding Random Effects

So far, the model has the structure of a traditional linear regression model (Gelman & Hill, 2006). However, mixed-effects models differ from regression models in that they include both fixed (i.e., predictors) and random effects, hence the name mixed effects (Bates, Mächler, Bolker, & Walker, 2015). A key assumption of regression analyses is that the error terms, the residual between each data point and the regression line, contain no systematic or patterned variations. In other words, because the error terms are random noise in the data, knowing how much one data point varies from the regression line would not give us an idea of how much another data point varies from the regression line (Bates et al., 2015). However, this is rarely the case when we obtain multiple within-subject responses because a participant who scores higher than the other participants in one response will likely score higher than the other participants in another response. To illustrate, suppose that each participant attempted to learn and remember words that varied from two to four syllables long. In this case, it is likely that there would be a main effect across participants in that increasing the number of syllables would lead to a decrease in the probability of retrieval (see Figure 2). Data from some participants would likely be close to the regression line and would have small residuals, but data from other participants would likely be further from the regression line and would have large residuals. Overall, it is more likely that a participant's data points will be more similar to each other in their residuals than they will be to data points from other participants (Baayen et al., 2008; Cnaan, Laird, & Slasor, 1997; Coupé, 2018; Goldenberg & Johnson, 2015).

In addition to similarities in residuals within participant, responses across participants to the same item will likely have similar residuals from the regression line (Judd,

Figure 2. Simulated data of three participants' ability to retrieve a word based on the number of syllables per word. A regression line is fitted to the data as a whole.

Westfall, & Kenny, 2012). If we consider our example from Figure 2, we see variations in responses to items in that items with fewer syllables have a higher probability of being retrieved. However, there may be some systematic variation in responses to items that we are not taking into account in the model. For example, we may find that participants have about an 80% probability of retrieving the two-syllable word, *keenit*. Thus, data points representing responses to this word will be close to the regression line. However, there may be another two-syllable word, *gramer*, that most participants struggle to retrieve. Responses to this word may be around 40% probability, and thus data points representing responses to this word will have much larger residuals from the regression line. In the ANOVA, I dealt with the variability of item residuals by making my outcome variable the total number of items correct for each participant. However, by doing this, I lost critical information, namely, how a participant's response to an item at the short-term test affects the probability of a correct response to that item at the long-term test.

I can control for this inter-subject and inter-item variability by including random effects for subjects and items in the model. Including random intercepts for subjects allows me to account for the fact that some participants are better at retrieving words than others. As you can see in the left panel of Figure 3, I fit a different regression line to each participant's data with different intercepts with the y -axis.

This provides a better fit to the data. In a similar manner, if I found systematic variability across responses to different items, such as all participants struggling to retrieve the word *gramer*, I could account for this variability by including random intercepts for items. Through including random intercepts, I can model more precisely the relationship between each predictor variable (e.g., number of syllables) and the outcome variable, which in turn decreases the likelihood of Type I and Type II errors (Baayen et al., 2008; Gries, 2015; Gueorguieva & Krystal, 2004; Quené & van den Bergh, 2008). Type I errors include reporting that a significant relationship exists between variables, when in fact it does not. Type II errors include reporting that a significant relationship between variables does not exist, when in fact it does (Johnson, 2014).

In addition to random intercepts, including random slopes in the model can help me determine more precisely how predictor variables relate to the outcome variable. As illustrated in the left panel of Figure 3, some participants were better at retrieving words than others. However, the probability of retrieving words decreased to a similar degree across participants as the number of syllables increased. However, suppose that the probability of retrieving words decreased to a different degree across participants as the number of syllables increased, as is shown in the right panel of Figure 3. By including random slopes in the model,

Figure 3. The panel on the left represents the same data from Figure 2, but with different intercepts for each participant. The panel on the right represents simulated data with different intercepts and different slopes for each participant.

I can better capture the nature of the relationship between the probability of retrieving a word and the number of syllables per word. Storkel et al. (2013) and Tomblin et al. (2015) provide examples of analyses that include both random intercepts and slopes.

Random intercepts and slopes for participants and items are commonly included in mixed-effects models in social science research. However, through mixed-effects modeling, we can control for other sources of systematic variation. For example, responses of children from one school district may be more similar to each other than responses of children from another school district. This systematic variation in the data can hamper our ability to understand the relationships between predictors, such as a specific reading intervention, and outcomes, such as changes in reading comprehension. Through mixed-effects modeling, we can include school district as a random effect to control for this variation. Similarly, we can control for other sources of systematic variation such as the effect of different clinicians implementing the same intervention (Fielding & Goldstein, 2006). We can even nest these sources of variation within each other in a hierarchical structure. For example, we can nest the data of children within classroom and also nest data from the individual classrooms within schools (Fielding & Goldstein, 2006; Gries, 2015). Within the hearing sciences, we can even nest data obtained from two

different ears within the same child (McCreery et al., 2016). This ability to nest data is why mixed-effects modeling is sometimes called *multilevel* or *hierarchical modeling* and can provide a powerful tool to account for variation in the data when identifying relationships between predictors and outcomes.

Fitting the Model

Let us return to the example data set of children's performance in the word learning task based on their age, vocabulary score, week of testing, and response at the short-term test (Gordon et al., 2016). To analyze these data, I conducted a mixed-effects logistic regression in an R environment using the lme4 package.¹ R is a free software program used by researchers to conduct a variety of statistical analyses (R Development Core Team, 2007). Learning how to analyze data in R can take more practice than using other common programs such as SPSS (see <http://cran.r-project.org/other-docs.html> for information about how to use R). In contrast to SPSS, R does not include drop-down menus that allow the user to select

¹This mixed-model analysis was conducted by Dr. Maura Curran in the original publication (Gordon et al., 2016), but I reran the analysis for the current article.

between various analyses. However, an advantage to using R is increased scientific rigor and transparency. Researchers can save the code they use to analyze data, providing a transparent record of the details of their analyses that can be revisited by the researcher themselves and shared with others. Within SPSS, once analyses are run, the details of those analyses are lost (McElreath, 2016). This can contribute to errors in reporting and limitations in sharing detailed information about analyses with others.

There are differing opinions on how to find the model that best fits the data when conducting mixed-effects models. However, I followed the recommendations of Crawley (2012) and Jaeger (2008) in that I first determined the minimal random effects structure that best supported model fit. In this case, eliminating either random intercepts for participants or items significantly reduced model fit, so I retained both. Including random slopes can improve model fit when there are multiple responses from each participant at different levels of a continuous variable. However, random slopes are not needed in the current model because the data include only one continuous variable, PPVT score, and each participant has only one score. After I determined that retaining random intercepts for subjects and items improved model fit, I then determined the fixed effects structure that best supported model fit. To do so, I fit the maximal model with all possible main effects and interactions between the predictor variables. The model that I have shown thus far only includes main effects and not interactions. The full model is not included here, because it is quite long, but see Appendix A for an example of a full model with all possible main effects and interactions.

I eliminated the main effect or interaction with the smallest coefficient and compared the simplified model to the maximal model to assess whether this change reduced model fit. I continued this process until the model converged, and I identified the minimal model that best supported model fit. The final model in this case included fixed effects for short-term response, week, age, and PPVT score and an interaction between age and short-term response (see Table 3). Thus, the final model has the structure shown below. Random effects are not shown in the model, but know that for each participant's response to each item,

we would have a random intercept for participant and a random intercept for item (see Locker et al., 2007, for an example).

$$\ln(p/1-p) = .80 + (.47 \times \text{Short-Term Response}) + (-0.97 \times \text{Week}) + (-0.61 \times \text{Age}) + (0.01 \times \text{PPVT}) + (1.18 \times \text{Age} \times \text{Short-Term Response}) + e \quad (7)$$

What can I conclude from this analysis about how individual factors affect learner outcomes? Similar to the ANOVA, I found a significant effect of week, $\beta = -.97$, $z = -3.78$, $p < .01$, in that children performed better during Week 1 than Week 2. The coefficient is negative in this case because performance during Week 1, which was coded as 0, was better than performance during Week 2, which was coded as 1. I also found a significant effect for raw PPVT score, $\beta = .01$, $z = 2.04$, $p < .04$. This result provides a more nuanced understanding of the effect of PPVT score on word retrieval than the ANOVAs. In this case, the coefficient for PPVT score is 0.01, which tells me that an increase of 1 in raw PPVT score would result in an increase of 0.01 in the log odds of a correct response at the long-term test. See Appendix B for details about how the coefficient of 0.01 for PPVT score relates to changes in the probability of a correct response.

In addition to these main effects, I found a significant Age $\times$ Short-Term Response interaction. To investigate this interaction, I can use the final model to estimate the probability of a correct response for each age group given either a correct or incorrect response at the short-term test. For example, to calculate the probability of correct response by a 3-year-old child at the long-term test who responded incorrectly at the short-term test, I would use the following formula:

$$\ln(p/1-p) = .80 + (.47 \times 0) + (-.97 \times 0.5) + (-0.61 \times 0) + (0.01 \times 0) + (1.18 \times 0 \times 0) \quad (8)$$

I entered 0 for the short-term response to reflect an incorrect response. I entered 0.5 for week to provide an average estimate across the two weeks; recall that Week 1 was coded as 0 and Week 2 was coded as 1. I entered 0 for age to reflect the younger age group and 0 as the PPVT score to calculate the response of a child with a mean vocabulary score. The log odds in this case is .315. If $\ln(p/1-p) = .315$, I can solve for p , using the formula below:

$$p = 1 / (1 + e^{-(.315)}) \quad (9)$$

Thus, the probability of a correct response under these circumstances is .58. In a similar way, I can calculate the probability of a correct response for each age group based on the short-term response, which reveals the nature of the age by short-term response interaction (see Table 4). For 4- and 5-year-old children, the probability of a correct

Table 3. Final mixed-effects model predicting the probability of a correct response in the long-term test.

Predictor	Estimate	SE	z	Pr(> z)
Intercept	0.80	0.39	2.06	0.04
Week ^a	-0.97	0.26	-3.78	< 0.001
PPVT raw score	0.01	0.01	2.04	0.04
Age ^b	-0.61	0.51	-1.21	0.23
Short-term response ^c	0.47	0.36	1.32	0.19
Age $\times$ Short-Term Response	1.18	0.53	2.22	0.03

Note. PPVT = Peabody Picture Vocabulary Test.

^aReference group is "Week 1." ^bReference group is "3-year-olds."

^cReference group is "incorrect."

Table 4. Probability of a correct response at the long-term test.

Short-term response	3-year-olds		4- and 5-year-olds	
	Log odds	Probability	Log odds	Probability
Short-term incorrect	0.315	.58	-0.295	.43
Short-term correct	0.785	.69	1.355	.79

response to the long-term test changed greatly based on their response at the short-term test. Specifically, if they responded incorrectly to an item at the short-term test, they had a .43 probability of responding correctly to that item at the long-term test. In contrast, if they responded correctly to an item at the short-term test, they had a .79 probability of responding correctly to that item at the long-term test. The probability of a correct response at the long-term test for 3-year-old children changed to a much lesser degree based on their response to the short-term test, .58 for an incorrect response to .69 for a correct response.

Comparison Between the ANOVA and Mixed-Effects Model Analyses

You may wonder why I got different results when I conducted the repeated-measures ANOVA as opposed to the mixed-effects model analysis. To establish an understanding of how these analyses are related, it is first important to know that repeated-measures ANOVA is a special case of regression analyses (Kleinbaum, Kupper, Nizam, & Rosenberg, 2013) in which all predictor variables are represented categorically (see Appendix A for more details). Regression analyses, as a broader category, can include either categorical predictors, continuous predictors, or both. Another key difference between ANOVA and regression analyses is what we typically use as the comparison group to establish significance (Rutherford, 2011). Although the comparison coding options can be changed in most statistical software programs, most users use the default options when running analyses. A standard ANOVA uses effect (i.e., sum) coding in which we ask whether the mean for each group is significantly different from the grand mean across groups. In contrast, standard regression analyses use reference (i.e., dummy, treatment) coding in which we ask whether the mean for each group differs significantly from the group that we designate as the reference group. Critically, through either comparison, we ask whether changes in the predictor variables affect the outcome variable in a significant way.

As already discussed, mixed-effects models are similar to regression analyses, but mixed-effects models include random effects to better control for systematic variation across the residuals. The repeated-measures ANOVA actually does include one random effect, a random intercept for participant, which is the repeated-measures component of this analysis. However, through a mixed-effects analysis, researchers can include random intercepts and slopes for a variety of sources of variation such as participant, items, schools, instructors, and so forth. An additional difference

between regression and mixed-effects models is that coefficients are calculated via the ordinary least squares approach when conducting regressions and via the maximum likelihood approach when conducting mixed-effects models (Baayen et al., 2008). What is important to know is that if the data are balanced, in that there are equal numbers of participants in each group, normally distributed, and there is little systematic variation in the residuals, both regression and mixed-effects models will yield very similar results (Misangry et al., 2006). However, if these assumptions are violated, these analyses can yield different results.

What does this mean for researchers, clinicians, and educators? As mentioned previously, mixed-effects modeling can help researchers avoid Type I and Type II errors, which can have some important real-world consequences for clinical and education practice. In a classic example, Bennett (1976) found that children taught with a formal style of instruction showed greater improvements in reading than those who were not. However, in the original analyses, the researchers did not take into account the systematic variation in the residuals from children who were taught by the same teacher or were from the same school. When these factors were taken into account in follow-up analyses, the significance of this effect disappeared (Aitkin, Anderson, & Hinde, 1981).

Additionally, beyond knowing whether a predictor variable is significantly related to an outcome variable, mixed-effects modeling allows researchers to gain a more accurate estimate of how each predictor variable is related to the outcome variable. For example, let us again consider the final model from the example.

$$\ln(p/1-p) = .80 + (.47 \times \text{Short-Term Response}) + (-0.97 \times \text{Week}) + (-0.61 \times \text{Age}) + (0.01 \times \text{PPVT}) + (1.18 \times \text{Age} \times \text{Short-Term Response}) + e \quad (10)$$

Based on this model, I know that a change from Week 1 to Week 2, with a coefficient of -0.97 , has a much larger effect on the probability of a correct response than an increase of 1 in raw PPVT score, with a coefficient of 0.01 . When coefficients are clearly reported, practitioners can estimate how a specific client with a given vocabulary score and a given age would likely perform in similar tasks (see Appendix B). Critically, mixed-effects models allow researchers to determine more accurate coefficients, providing a more accurate estimate of the true relationships between variables. Information about these relationships can have real-world consequences as practitioners translate research findings to practice.

Three Additional Benefits of Mixed-Effects Models

There are some additional advantages of mixed-effects models in answering questions about learning and memory. Specifically, mixed-effects models provide the means to account for missing data, assess the contribution of variations

in delay intervals, and model nonlinear relationships between factors.

Missing Data

Missing data is a common problem in longitudinal work, and longitudinal work is essential to understand learning and memory processes. The more sessions researchers add to a study and the longer the delay between sessions, the more likely it is that participants will not complete all sessions. Excluding all of the data from participants with missing data can create biases in the data set in that these participants are likely to differ in systematic ways from those who complete all sessions (Cnaan et al., 1997; Gueorguieva & Krystal, 2004). Additionally, despite best efforts to reduce experimenter error, it is always a part of data collection. For research studies that include only one session for each participant, it is often not a great loss to exclude all of their data due to experimenter error. However, it can be a great loss of time and effort to exclude data in a longitudinal study, especially when the error occurs near the end of the study. Furthermore, for researchers who study special populations, excluding all data from a participant because of experimenter error greatly slows the process of research.

A strength of mixed-effects models is that data from a given participant with missing data can be retained in the analyses (Collins, 2006). Through mixed-effects modeling, we are estimating the effects of predictor variables on the outcome variable (i.e., fixed effects) while taking into account systematic patterns in the residuals through random effects (i.e., subject and item random effects). Thus, we can account for the missing response of a participant to a particular item by using the information that we have on that participant's responses to other items as well as the information that we have from all of the other participants' responses to that item (Baayen et al., 2008; Cnaan et al., 1997). Retaining all of the data provides more statistical power to identify relationships between predictors and outcome variables. Critically, the ability to retain all of the data allows researchers to provide higher quality information to educators and clinicians.

Variations in Delays

Another major challenge of longitudinal research is that participants are likely to complete follow-up sessions after various delays. For example, we may be interested in children's memory for taught items after a 48-hour delay. However, it is unlikely that each child will complete the long-term test exactly 48 hours after training. Typically, we either ignore these variations in delays or exclude data from sessions that did not conform with the schedule. However, variations in delays across participants could provide valuable information about the effect of the delay on learning and memory, but we need statistical tools that allow us to assess these effects. Below, I include an example of how this can be done.

If you recall from the example data set (Gordon et al., 2016), children were trained and tested on two sets of word-object pairs across two weeks. After children completed this training, we were interested in their ability to identify the correct word form for each trained object after a much longer delay. Nineteen children completed the very long-term test, six months to one year after training,² in which they were asked to identify the name of each object. Thus, these 19 children responded to each item at three time points: at the end of training (short-term delay), after a two-day delay (long-term delay), and six months to one year later (very long-term delay). The model to predict the probability of a correct response at the very long-term delay is as follows:

$$\ln(p/1-p) = \beta_0 + (\beta_1 \times \text{Short-Term Response}) + (\beta_2 \times \text{Week}) + (\beta_3 \times \text{Age}) + (\beta_4 \times \text{PPVT}) + (\beta_5 \times \text{Long-Term Response}) + (\beta_6 \times \text{Delay in Days}) + e \quad (11)$$

Once again, only main effects and not interactions are shown here, but we did include all interactions in the maximal model.³ For this model, we standardized delay in days via a conversion to group *z* scores. For the random effects structure, we tested model fit with random intercepts for subjects and items. Excluding either of these factors reduced model fit, so both were retained. We included a random slope for participant in the model because we have multiple responses from each participant along a continuous variable, time. Eliminating this random effect reduced model fit, so it was retained.

We determined that the minimal fixed effect structure that supported model fit included main effects for age, short-term response, and delay in days and an interaction between short-term response and age (see Table 5).

$$\ln(p/1-p) = \beta_0 + (\beta_1 \times \text{Short-Term Response}) + (\beta_2 \times \text{Age}) + (\beta_3 \times \text{Delay in Days}) + (\beta_4 \times \text{Age} \times \text{Short-Term Response}) + e \quad (12)$$

An investigation of the Short-Term Response $\times$ Age interaction revealed that a correct response at the short-term test substantially changed the probability of a correct response at the very long-term test for the older children, but not for the younger children. Interestingly, we found a trend of delay in days in that the probability of a correct response decreased as the retention interval increased.

²Children initially participated between September 2014 and May 2014, a range of nine months. For the very long-term test, they were asked to return between September 2015 and December 2015. Thus, there were variations in how much time had passed since training for each child; the range was five months 26 days to one year 24 days.

³Dr. Maura Curran conducted all of these analyses (Gordon et al., 2016).

Table 5. Final mixed-effects model predicting response at the very long-term test.

Predictor	Estimate	SE	z	Pr(> z)
Intercept	−0.50	0.33	−1.52	0.13
Short-term response ^a	0.48	0.37	1.31	0.19
Age ^b	−1.01	0.71	−1.42	0.16
Delay (standardized) ^c	−0.29	0.15	−1.92	0.06
Age × Short-Term Response	1.67	0.80	2.08	0.04

^aReference group is “incorrect.” ^bReference group is “3-year-olds.”
^cDelay in days from the long-term test to the very long-term test is standardized.

The advantage of using mixed-effects model analyses to determine the effect of delay on performance in learning and memory research is that we can include the exact duration of the delay in the model. We cannot do this in an ANOVA, in which data are analyzed categorically. For example, in an ANOVA, we would compare whether children’s responses at the short-term test differed significantly from their responses at the long-term and very long-term tests. Identifying the effect of delay in a continuous manner is a powerful tool to increase the effectiveness and efficiency of educational and clinical practice. Understanding the time scale at which memories for taught items fade can help practitioners identify how often trained material should be reviewed to maintain what was taught. In this way, practitioners can avoid wasting time reviewing material more often than is needed or not reviewing material often enough.

Modeling Nonlinear Relationships

Many types of learning show nonlinear rates of change (De Bot, Lowie, & Verspoor, 2007). For example, individuals may demonstrate periods of rapid learning, but then the learning rate slows or levels off (see Figure 4; see C. C. Dunn et al., 2014, for an example). Similarly, forgetting rates are likely to follow nonlinear trajectories as demonstrated by the classic forgetting curve in which information is forgotten rapidly at first but then the forgetting rate slows until it reaches an asymptote (Murre & Dros, 2015). For example, if trained information is not reviewed, an individual may forget most of the information over the course of a week but retain 10% of the information over several months. Asymptotic models can be fitted to data through mixed-effects analyses with the following model:

$$y = a - be^{-cx} \quad (13)$$

In this case, a represents the level of the asymptote, b represents the difference between the starting value and the asymptote, c represents the rate of change from the intercept to the asymptote, x represents a continuous variable such as time (e.g., different training sessions or delay intervals), and e represents the residuals. The ability to model nonlinear relationships provides us with a powerful tool to

determine how predictor variables differentially contribute to learning and forgetting rates. We can determine individual or group differences in baseline performance (e.g., the intercept), differences in asymptotes (e.g., the maximum amount of information that was learned), and differences in learning rates. Additionally, we can assess how differences in training material, such as the number and type of items trained or training strategy, affect the intercept, asymptote, and rate of learning and forgetting. Alt and Spaulding (2011), Cnaan et al. (1997), Dunn et al. (2014), and Storkel et al. (2013) provide examples of data that demonstrated a nonlinear relationship between predictors and outcomes.

The ability to fit nonlinear models to data is a powerful tool to improve clinical and educational practice. Understanding similarities and differences across populations in baseline performance, maximum performance achieved, and changes in learning and forgetting rates will help clinicians tailor their interventions to the needs of their clients. For example, understanding which teaching strategy leads to the most information retained long-term (i.e., highest asymptote) would help educators maximize the effectiveness of their teaching. Similarly, understanding the expected learning and forgetting trajectories of a child with specific characteristics would help a clinician assess whether the client is making adequate progress with a given treatment option. Overall, to maximally support learning and memory, it is essential that we understand how learner factors, training materials, and training strategies alter the characteristics of nonlinear trajectories.

Example 2: How Characteristics of Training Affect Learner Outcomes

As a final example, I analyzed some pilot data collected as part of a larger project to illustrate how we can use mixed-effects models to identify effective and efficient training strategies. In this case, four preschool-age children with typical development were trained on a set of nine word–object pairs. During each training day, they were first given a free recall test in which they were shown each object and asked, “What is this one called?”. They were then trained on all words that they missed in this free recall test. At the end of training for that day, they were given a recognition test in which they were presented with the target word form (e.g., duver), another form linked to another object during training (e.g., tiln), and minimal pairs of both forms (e.g., What is this one called? Is it a duver, a duler, a tiln, or a miln?). They were also given a free recall test (e.g., What is this one called?) and a cued recall test (i.e., What is this one called? It starts with du...) for all nine word–object pairs (see Table 6). Note that the first training day differed from the others in that each object was explicitly named before children were tested, and pairs were trained and tested in three sets of three. For the purposes of this example, it is important to know that each training day included a free recall test at the beginning of the session and a recognition

Figure 4. Simulated learning and forgetting rates. For the learning rate in the left panel, the asymptote is $a = .8$, the difference between the starting value and the asymptote is $b = .5$, and the rate of change is $c = .5$. For the forgetting rate in the right panel, the asymptote is $a = .2$, the difference between the starting value and the asymptote is $b = .5$, and the rate of change is $c = .5$.

test, a free recall test, and a cued recall test at the end of the session.

Children continued training until they reached criterion⁴ on all nine word–object pairs or completed a total of six training days. Two days after the last training day, they were tested on all pairs via a free recall test. Three of the children took three, three, and four training days to reach criterion. One child completed six training days but only responded correctly to seven of nine items on the last training day. Although the three children who reached criterion responded correctly to all items on the last training day and thus technically learned each word–object pair, their pattern of responses to each item varied throughout training. Specifically, they responded correctly to some items on the first or second training day and continued to respond correctly to those items throughout training. In contrast, they did not respond correctly to other items until

the last training day. For the child who did not reach criterion, he responded correctly to all items at some point during training, demonstrating some learning of all items. However, he did not reach criterion because he did not respond correctly to all items at the end of any one training day.

Given this variable pattern of responses, I can assess whether responses to each item during training predicted the probability of successful retrieval after a two-day delay. To do this, I calculated the percentage of times each child responded correctly to each item in each test across the training days and entered these values into a mixed-effects model. I used percentage correct instead of raw values because the total number of training days varied across children. Thus, the fixed effects of the maximal model included percentage correct in the free recall test given at the beginning of each session, percentage correct to the free recall test given at the end of each session, and percentage correct to the cued recall test given at the end of each session. I did not include responses to the recognition test as a fixed effect because all children performed near ceiling, and thus performance was unlikely to be predictive of retrieval of items after the two-day delay. Random effects included random intercepts for items. Given that I only

⁴To reach criterion, they were required to demonstrate a memory for all nine forms and the link between the form and the object at the end of a training day. Thus, they had to either respond correctly in the free recall test or respond correctly in both the recognition and cued recall tests to all nine items at the end of a training day to reach criterion.

Table 6. Methods for the second example.

Training day 1	Training day 2	Training day 3	Training day X...	Test after two-day delay
Free recall Training * * * * * * * * *	Free recall Training * * * * * *	Free recall Training * * *	Free recall Training *	Free recall test
Recognition Free recall Cued recall	Recognition Free recall Cued recall	Recognition Free recall Cued recall	Recognition Free recall Cued recall	

Note. All items were tested at the beginning and end of each session. Training each day only included items missed in the free recall test at the beginning of the session (represented by number of asterisks [*]).

had data from four participants, the best way to conduct the model would likely be to include participant as a fixed effect. However, to simplify this example for illustrative purposes, I elected to not include participant as a fixed or random effect. Middleton, Schwartz, Rawson, and Garvey (2015) provide an excellent example of how to conduct a mixed-effects model with participant as a fixed effect.

I conducted a mixed-effects logistic regression in an R environment using the lme4 package. The minimal fixed effects model that best supported data fit included a fixed effect for response to the free recall test at the beginning of each session. Thus, the final model was

$$\ln(p/1-p) = -0.63 + (5.57 \times \text{FreeRecallBeginning}) + e \quad (14)$$

Using this model, I can calculate the probability that a child will retrieve an item after the two-day delay based on the percentage of times they responded correctly to that item in the free recall test at the beginning of each session. For example, if a child successfully retrieved an item 65% of the time in the free recall test at the beginning of each session, the probability of a correct response to that item after the two-day delay would be .95, which is demonstrated below.

$$\ln(p/1-p) = -0.63 + (5.57 \times .65) \quad (15)$$

$$\ln(p/1-p) = 2.99 \quad (16)$$

$$p = .95 \quad (17)$$

Using this model, I can calculate how variations in children's responses to items during training are related to the probability of retrieving that item after the two-day delay (see Figure 5). Note that, for items that were missed often during training, a small change in the percentage correct during training led to a large change in the probability of a correct response after the two-day delay. For example, a change from 15% to 25% correct during training led to a

change in the probability of retrieval from .55 to .68. However, for items that were responded to correctly 65% of the time or greater, changes in the percentage correct during training did not lead to much gain in the probability of a correct response after the two-day delay. For example, a change from 65% to 75% only led to a change in the probability of retrieval from .95 to .97.

Given that this model is based on the data of four children, the predictive power is limited. However, hopefully you can see the usefulness of such models to maximize the efficiency of training to support long-term retention. For clinicians, knowing the ideal level of performance their clients should reach on targeted material before they discontinue training would improve the efficiency and effectiveness of interventions (Moore, 2018). Through using mixed-effects models, researchers can identify the level of performance that should be reached before training is discontinued to maximize long-term retention given individual characteristics, training material, and training strategies. In this way, researchers can improve the information they provide to practitioners about how to best support learning and memory in others.

Conclusion

Learning and retention are complicated processes. However, understanding how to support both is vital to achieve educational and clinical goals. Mixed-effects modeling provides a powerful tool to gain this understanding. Namely, through using this statistical technique, we can better determine how characteristics of individuals, training materials, and training strategies affect the likelihood that information will be learned and retained. Additionally, through mixed-effects modeling we can account for missing data, determine the effect of various delay intervals, and model nonlinear relationships. As is true for any research in the social sciences, the predictive power of research findings

Figure 5. Probability of a correct response after a two-day delay based on responses during training.

will always be limited by the characteristics of the participants, the sample size, and the sensitivity of the included measures to assess the variables of interest. However, researchers can improve their ability to discover accurate information about learning and retention through carefully planned methodologies and statistical techniques that make the best use of all the data that they have. Additionally, they can improve the ability of practitioners to apply findings to real-world settings through including clear and detailed information about the statistical nature of the relationships between predictor and outcome variables. In turn, practitioners can improve the application of research findings to practice by enhancing their statistical literacy. Through these efforts, we can achieve the joint goal of improving outcomes for the individuals we serve.

Acknowledgments

The author's data and findings presented in this publication were supported by funding by organizations within the National Institutes of Health including the National Institute on Deafness and Other Communication Disorders Grant F32DC013704-03 (principal investigator: Katherine R. Gordon) and the National Institute of General Medical Sciences Grant P20GM109023 (principal investigator: Lori J. Leibold). The author would like to thank the reviewers for the suggestion to make this article more accessible and directly relevant to clinicians and educators. The author would

also like to thank Karla McGregor and Nancy Ohlmann for suggestions on how to tailor the article for the clinical audience and Jacob Oleson for answering questions about the nuanced differences between statistical analyses.

References

- Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, 144*(4), 419–461.
- Alt, M., Meyers, C., & Ancharski, A. (2012). Using principles of learning to inform language therapy design for children with specific language impairment. *International Journal of Language & Communication Disorders, 47*(5), 487–498. <https://doi.org/10.1111/j.1460-6984.2012.00169.x>
- Alt, M., & Spaulding, T. (2011). The effect of time on word learning: An examination of decay of the memory trace and vocal rehearsal in children with and without specific language impairment. *Journal of Communication Disorders, 44*(6), 640–654. <https://doi.org/10.1016/j.jcomdis.2011.07.001>
- American Speech-Language-Hearing Association. (2010). *Roles and responsibilities of speech-language pathologists in schools*. Retrieved from <https://www.asha.org/policy/PI2010-00317/>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).
- Bennett, N. (1976). *Teaching styles and pupil progress*. London, United Kingdom: Open Books.
- Cnaan, A., Laird, N., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16, 2349–2380. [https://doi.org/10.1002/\(SICI\)1097-0258\(19971030\)16:20<2349::AID-SIM667>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E)
- Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, 57(1), 505–528. <https://doi.org/10.1146/annurev.psych.57.102904.190146>
- Common Core State Standards Initiative. (2010). *Common core state standards for English, language arts and literacy in history/social studies, science and technical subjects*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.
- Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, 9, 1–21. <https://doi.org/10.3389/fpsyg.2018.00513>
- Crawley, M. J. (2012). *The R book*. Chichester, United Kingdom: Wiley.
- De Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1), 7–21. <https://doi.org/10.1017/S1366728906002732>
- Dunn, C. C., Walker, E. A., Oleson, J., Kenworthy, M., Voorst, T. V., Tomblin, J. B., ... Gantz, B. J. (2014). Longitudinal speech perception and language performance in pediatric cochlear implant users: The effect of age at implantation. *Ear and Hearing*, 35(2), 148–160. <https://doi.org/10.1097/AUD.0b013e3182a4a8f0>
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test—Fourth Edition (PPVT-4)*. Bloomington, MN: NCS Pearson.
- Edgar, D. L., & Rosa-Lugo, L. I. (2007). The critical shortage of speech-language pathologists in the public school setting: Features of the work environment that affect recruitment and retention. *Language, Speech, and Hearing Services in Schools*, 38(1), 31–46.
- Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 129 Stat. 1802 (2015).
- Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Research Report no. 791). Birmingham, United Kingdom: University of Birmingham.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.
- Goldenberg, E. R., & Johnson, S. P. (2015). Category generalization in a new context: The role of visual attention. *Infant Behavior & Development*, 38, 49–56. <https://doi.org/10.1016/j.infbeh.2014.12.001>
- Gordon, K. R., McGregor, K. K., Waldier, B., Curran, M. K., Gomez, R. L., & Samuelson, L. K. (2016). Preschool children's memory for word forms remains stable over several days, but gradually decreases after 6 months. *Frontiers in Psychology*, 7, 1439. <https://doi.org/10.3389/fpsyg.2016.01439>
- Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125. <https://doi.org/10.3366/cor.2015.0068>
- Georgieva, R., & Krystal, J. H. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of General Psychiatry*, 61(3), 310–317. <https://doi.org/10.1001/archpsyc.61.3.310>
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Johnson, D. E. (2014). *Progress in regression: Why natural language data calls for mixed-effects models*. Lancaster University. Retrieved from http://danielezrajohnson.com/johnson_2014.pdf
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69.
- Justice, L. M., Logan, J., Schmitt, M. B., & Jiang, H. (2016). Designing effective speech-language interventions for children in the public schools: Leverage the spacing effect. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 85–91. <https://doi.org/10.1177/2372732215624705>
- Katz, L. A., Maag, A., Fallon, K. A., Blenkarn, K., & Smith, M. K. (2010). What makes a caseload (un)manageable? School-based speech-language pathologists speak. *Language, Speech, and Hearing Services in Schools*, 41(2), 139–151.
- Kelley, E., Leary, E., & Goldstein, H. (2018). Predicting response to treatment in a Tier 2 supplemental vocabulary intervention. *Journal of Speech, Language, and Hearing Research*, 61(1), 94–103.
- Kleinbaum, D., Kupper, L., Nizam, A., & Rosenberg, E. (2013). *Applied regression analysis and other multivariable methods* (5th ed.). Boston, MA: Cengage Learning.
- Law, J., Tomblin, J. B., & Zhang, X. (2008). Characterizing the growth trajectories of language-impaired children between 7 and 11 years of age. *Journal of Speech, Language, and Hearing Research*, 51(3), 739–749.
- Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, 39(4), 723–730. <https://doi.org/10.3758/BF03192962>
- Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T., & Saxena, S. (2011). Prevalence of intellectual disability: A meta-analysis of population-based studies. *Research in Developmental Disabilities*, 32(2), 419–436.
- McCreery, R. W., Walker, E. A., Spratford, M., Kirby, B., Oleson, J., & Brennan, M. (2016). Stability of audiometric thresholds for children with hearing aids applying the AAA pediatric amplification guidelines: Implications for safety. *Journal of the American Academy of Audiology*, 27(3), 252–263.
- McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.
- McGregor, K. K., Gordon, K., Eden, N., Arbisi-Kelm, T., & Oleson, J. (2017). Encoding deficits impede word learning and memory in adults with developmental language disorders. *Journal of Speech, Language, and Hearing Research*, 60(10), 2891–2905.
- Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(4), 1253–1261. <https://doi.org/10.1037/xlm0000091>

- Mirman, D.** (2016). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.
- Misangyi, V., LePine, J., Algina, J., & Goeddeke Jr., F.** (2006). The adequacy of repeated-measures regression for multilevel research: Comparisons with repeated-measures ANOVA, multivariate repeated-measures ANOVA, and multilevel modeling across various multilevel research designs. *Organizational Research Methods*, 9(1), 5–28. <https://doi.org/10.1177/1094428105283190>
- Moore, R.** (2018). Beyond 80-percent accuracy: Consider alternative objective criteria in writing your treatment goals. *The ASHA Leader*, 23(5), 6–7.
- Morrison, G. S., & Kondaurova, M. V.** (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America*, 126(5), 2159–2162. <https://doi.org/10.1121/1.3216917>
- Murre, J. M. J., & Dros, J.** (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One*, 10(7), e0120644. <https://doi.org/10.1371/journal.pone.0120644>
- Nunes, L. D., & Karpicke, J. D.** (2015). Retrieval-based learning: Research at the interface between cognitive science and education. In R. A. Scott & M. C. Buchmann (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781118900772.etrds0289>
- Perry, L. K., & Kucker, S. C.** (2019). The heterogeneity of word learning biases in late-talking children. *Journal of Speech, Language, and Hearing Research*, 62, 554–563. https://doi.org/10.1044/2019_JSLHR-L-ASTM-18-0234
- Powell, R. K.** (2018). Unique contributors to the curriculum: From research to practice for speech-language pathologists in schools. *Language, Speech, and Hearing Services in Schools*, 49(2), 140–147.
- Quené, H., & van den Bergh, H.** (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1–2), 103–121. <https://doi.org/10.1016/j.specom.2004.02.004>
- Quené, H., & van den Bergh, H.** (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4), 413–425. <https://doi.org/10.1016/j.jml.2008.02.002>
- R Development Core Team.** (2007). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org>
- Rice, M. L., & Hoffman, L.** (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2;6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*, 58(2), 345–359. https://doi.org/10.1044/2015_JSLHR-L-14-0150
- Rowland, C. A.** (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Rutherford, A.** (2011). *ANOVA and ANCOVA: A GLM approach*. Hoboken, NJ: Wiley.
- Storkel, H. L.** (2018). Implementing evidence-based practice: Selecting treatment words to boost phonological learning. *Language, Speech, and Hearing Services in Schools*, 49(3), 482–496. https://doi.org/10.1044/2017_LSHSS-17-0080
- Storkel, H. L., Bontempo, D. E., Aschenbrenner, A. J., Maekawa, J., & Lee, S.-Y.** (2013). The effect of incremental changes in phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Speech, Language, and Hearing Research*, 56(5), 1689–1700. [https://doi.org/10.1044/1092-4388\(2013\)12-0245](https://doi.org/10.1044/1092-4388(2013)12-0245)
- Storkel, H. L., Komesidou, R., Fleming, K. K., & Romine, R. S.** (2017). Interactive book reading to accelerate word learning by kindergarten children with specific language impairment: Identifying adequate progress and successful learning patterns. *Language, Speech, and Hearing Services in Schools*, 48(2), 108–124. https://doi.org/10.1044/2017_LSHSS-16-0058
- Tomblin, J. B., Harrison, M., Ambrose, S., Walker, E. A., Oleson, J., & Moeller, M. P.** (2015). Language outcomes in young children with mild to severe hearing loss. *Ear and Hearing*, 36, 76S–91S. <https://doi.org/10.1097/AUD.0000000000000219>
- Walker, E. A., Redfern, A., & Oleson, J.** (2019). Linear mixed-model analysis to examine longitudinal trajectories in vocabulary depth and breadth in children who are hard of hearing. *Journal of Speech, Language, and Hearing Research*, 62, 525–542. https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0250
- Waterhouse, L.** (2013). *Rethinking autism: Variation and complexity*. London, United Kingdom: Academic Press.

Appendix A

A Comparison Between ANOVA and Regression Analyses

I conducted a regression analysis to model the number of items that a child would get right at the long-term test in the same way that I ran the ANOVA. In this case, the predictor variables were delay (short-term = 0, long-term=1), week (Week 1 = 0, Week 2 = 1), age group (younger = 0, older = 1), and PPVT group (low vocabulary = 0, high vocabulary = 1). The outcome variable was the total number of items correct out of six. Thus, the regression model with all possible main effects and interactions has the following structure:

$$\begin{aligned}
 Y = & \beta_0 + (\beta_1 \times \text{Delay}) + (\beta_2 \times \text{Week}) + (\beta_3 \times \text{Age Group}) + (\beta_4 \times \text{PPVT Group}) + (\beta_5 \times \text{Delay} \times \text{Week}) \\
 & + (\beta_6 \times \text{Delay} \times \text{Age}) + (\beta_7 \times \text{Week} \times \text{Age}) + (\beta_8 \times \text{Delay} \times \text{PPVT}) + (\beta_9 \times \text{Week} \times \text{PPVT Group}) \\
 & + (\beta_{10} \times \text{Age} \times \text{PPVT Group}) + (\beta_{11} \times \text{Delay} \times \text{Week} \times \text{Age}) + (\beta_{12} \times \text{Delay} \times \text{Week} \times \text{PPVT Group}) \\
 & + (\beta_{13} \times \text{Delay} \times \text{Age} \times \text{PPVT Group}) + (\beta_{14} \times \text{Week} \times \text{Age} \times \text{PPVT Group}) + (\beta_{15} \times \text{Delay} \times \text{Week} \\
 & \times \text{Age} \times \text{PPVT Group}) + e
 \end{aligned} \quad (1)$$

To make the regression analysis comparable to the repeated-measures ANOVA, I included a random intercept for participant. After conducting the regression, I get the coefficients for the main effects and interactions listed in the estimates column of Table A1.

Using the coefficients, I can calculate the mean response of the older age group by entering the appropriate values in the model (see Column X of Table A1). I entered 0.5 for delay and week to provide an average response across weeks and delay intervals. I entered 1 for age to reflect the older age group and 0.5 for PPVT group to provide an average across the low and high PPVT groups. The value I entered for the interaction terms is simply the product of the values I entered for each main effect. I can now solve the entire equation by multiplying each estimate by the predictor value I entered and then summing all of the products (see Estimate $\times$ X column of Table A1). For this example, the mean response for the older age group is 4.281 items retrieved out of six. In a similar way, I can determine the mean response for the younger age group by entering the value of 0 for all the main effects and interactions that include age. Thus, the mean response for the younger age group is 3.50 items retrieved out of six.

The mean number of items retrieved for the younger and older groups that I calculated via the regression model is the exact means for the data I analyzed via the ANOVA. However, the results for the two analyses are different. For example, the regression analysis revealed only one significant effect, a significant effect of week, $t = -2.72$, $p < .01$. The ANOVA also revealed a significant effect of week, but the p value differed, $F(1, 28) = 31.18$, $p < .001$. In addition, the ANOVA revealed a significant Delay $\times$ Vocabulary Group interaction, $F(1, 28) = 5.17$, $p = .03$, although the regression analyses did not. These differing results are due to what I used as the comparison group in each analysis. For the ANOVA, I used effect (i.e., sum) coding in which I asked whether the mean for each group is significantly different from the grand mean across groups. For the regression analysis, I used reference (i.e., dummy, treatment) coding in which I asked whether the mean for each group differs significantly from the group that we designate as the reference group for each predictor. For example, the younger age group was the reference group for age, and through the analysis, I determined if being in the older age group significantly changed the number of correct responses in comparison to being in the younger age group. Thus, through each analysis, I was asking slightly different questions that lead to different p values. However, both analyses are using the same linear model to determine the relationship between the predictor variables and the outcome variable.

Table A1. Results from the regression analysis predicting number of items correct at the long-term test.

Predictor	Estimate	X	Estimate $\times$ X
Intercept	4.75	1	4.75
Delay ^a	-0.50	0.5	-0.25
Week ^b	-1.50	0.5	-0.75
Age group ^c	0.125	1	0.125
PPVT group ^d	-1.125	0.5	-0.563
Delay $\times$ Week	0.375	$0.5 \times 0.5 = 0.25$	0.094
Delay $\times$ Age	0.125	$0.5 \times 1 = 0.5$	0.063
Week $\times$ Age	0.375	$0.5 \times 1 = 0.5$	0.188
Delay $\times$ PPVT	0.50	$0.5 \times 0.5 = 0.25$	0.125
Week $\times$ PPVT	0.375	$0.5 \times 0.5 = 0.25$	0.094
Age $\times$ PPVT	1.50	$1 \times 0.5 = 0.5$	0.75
Delay $\times$ Week $\times$ Age	-1.00	$0.5 \times 0.5 \times 1 = 0.25$	-0.25
Delay $\times$ Week $\times$ PPVT	0	$0.5 \times 0.5 \times 0.5 = 0.125$	0
Delay $\times$ Age $\times$ PPVT	0.25	$0.5 \times 1 \times 0.5 = 0.25$	0.063
Week $\times$ Age $\times$ PPVT	-0.875	$0.5 \times 1 \times 0.5 = 0.25$	-0.219
Delay $\times$ Week $\times$ Age $\times$ PPVT	0.50	$0.5 \times 0.5 \times 1 \times 0.5 = 0.125$	0.063
		Sum	4.281

Note. PPVT = Peabody Picture Vocabulary Test.

Appendix B

How the Coefficients in the Model Relate to the Probability of a Correct Response

I conducted a mixed-effects logistic regression in an R environment using the lme4 package to predict the log odds of identifying the correct word form for each object at the long-term test. I included week of training (Week 1, Week 2), raw PPVT score, age (younger = 3-year-olds, older = 4- and 5-year-olds), and response to the short-term test (correct, incorrect) as predictors. The final model is as follows:

$$\ln(p/1-p) = .80 + (.47 \times \text{Short-Term Response}) + (-0.97 \times \text{Week}) + (-0.61 \times \text{Age}) + (0.01 \times \text{PPVT}) + (1.18 \times \text{Age} \times \text{Short-Term Response}) + e \quad (1)$$

To illustrate how the coefficient for each predictor relates to the probability of a correct response at the long-term test, let us consider the example of PPVT score. I will first calculate the log odds of a correct response for a child with a mean raw PPVT score, which if you recall is centered at 0.

$$\ln(p/1-p) = .80 + (.47 \times 0.5) + (-.97 \times 0.5) + (-0.61 \times 0.5) + (0.01 \times 0) + (1.18 \times 0.05 \times 0.05) + e \quad (2)$$

In this case, I entered 0.5 for short-term response, week, and age to provide an average estimate for these factors. Recall that, for the short-term response, incorrect = 0 and correct = 1; for week, Week 1 = 0 and Week 2 = 1; and for age group, 3-year-olds = 0 and 4- and 5-year-olds = 1. After entering these values, I see that the log odds of a correct response at the long-term test for a child with a mean PPVT score is 0.54.

If $\ln(p/1-p) = .54$, I can solve for p , using the formula below to determine the probability of a correct response under these circumstances, which in this case is .632.

$$p = 1 / (1 + e^{-(.54)}) \quad (3)$$

I can also calculate the log odds of a correct response for a raw PPVT score that is one item greater than the mean. Given that the coefficient for PPVT score is 0.01, an increase in a value of 1 in PPVT score should result in an increase of 0.01 of the log odds of a correct response as follows.

$$\ln(p/1-p) = .80 + (.47 \times 0.5) + (-.97 \times 0.5) + (-0.61 \times 0.5) + (0.01 \times 1) + (1.18 \times 0.05 \times 0.05) \quad (4)$$

In this case, the log odds of a correct response is 0.55, which reflects the 0.01 increase in log odds from the value of .54 that I found above. This can also be converted to the probability of a correct response at the long-term test, which is 0.634. I can use the model in a similar way to calculate the probability of a correct response at the long-term test of a child with a PPVT score, one standard deviation above the mean, one standard deviation below the mean, or any value I wish.