Modeling Longitudinal Outcomes: A Contrast of Two Methods.

Lohse, K. R.,[1,2], Jincheng Shen[3], & Kozlowski, A. J.,[4,5]

1. University of Utah; Department of Health, Kinesiology, and Recreation
2. University of Utah; Department of Physical Therapy and Athletic Training
3. University of Utah; Department of Population Health Sciences
4. Michigan State University – College of Human Medicine; Department of Epidemiology and Biostatistics
5. Mary Free Bed Rehabilitation Hospital; John F. Butzer Center for Research & Innovation

Corresponding Author:
Keith Lohse, PhD
e: rehabinformatics@gmail.com
p: 801-585-7226
250 S 1850 E, Rm 258
Salt Lake City, UT, 84112

Contents: 5 Tables, 4 Figures

Acknowledgments: The authors would like to thank Rhiannon Cowan, Brad Fawver, Brady De Couto, and Mark Williams for providing access to the data that formed the basis of our simulations. We would also like to thank two anonymous reviewers whose careful feedback helped us balance the complexity, clarity, and ultimately the utility of this manuscript.

Disclosure Statement: The authors received no funding specifically to pursue this work and have no conflicts of interest to declare. Please note that this self-archived post-print corresponds to the authors' copy prior to final peer-review, copy editing, and formatting. The final version of the accepted manuscript is available from the *Journal of Motor Learning and Development* © Human Kinetics.

Abstract

Background: Repeated measures analysis of variance (ANOVA) is frequently used to model longitudinal data but does not appropriately account for within-person correlations over time, does not explicitly model time, and cannot flexibly handle missing data. In contrast, mixed-effects regression addresses these limitations. In this commentary, we compare these two methods using openly available tools.

Methods: We emulated a real developmental study of elite skiers, tracking national rankings from 2011 to 2018. We constructed unconditional models of time (establishing the "pattern" of change), conditional models (identifying factors that affect change over time) and contrasted these models against comparable repeated measures ANOVAs.

Results: Mixed-effects regression allowed for linear and non-linear modeling of the skiers' longitudinal trajectories despite missing data. Missing data is still a concern in mixed-effects regression models, but in the present dataset missingness could be accounted for by skiers' ages, satisfying the missing at random assumption.

Discussion: Although ANOVA and mixed-effects regression are both suitable for time-series data, their applications differ. ANOVA will be most parsimonious when the research question focuses on group-level mean differences at arbitrary time points. However, mixed-effects regression is more suitable where time is inherently important to the outcome, and where individual differences are of interest.

Evaluating change over time is a common aspect of research in motor learning and development. These might be changes during life-span development (e.g., the emergence of fundamental motor skills), changes due to experience (e.g., learning the coordination to juggle through practice), or changes following illness and injury (e.g., recovery of motor functions following a stroke). There are numerous designs (e.g., cohort studies, longitudinal studies, time to event analyses) and analysis methods available for studying change over time. In this commentary, our contention is that motor learning and development researchers continue to be overly reliant on analysis of variance (ANOVA) approaches relative to other methods that are available.

To illustrate this point, we conducted a review of four motor behavior journals (*JMLD*, *Human Movement Science*, *Journal of Motor Behavior*, and *Motor Control*) over the last 5 years. Of the 2041 articles found on Google Scholar from January 1st 2014 to June 20th 2019, we found that 758 used descriptors for ANOVA-based methods (viz., "repeated-measures", "mixed-factor", "mixed-factorial"), whereas only 15 used descriptors for mixed-effect regression models (viz., "mixed-effect", "multilevel model" or "MLM", "hierarchical linear model" or "HLM"). Of the 15 articles that employed mixed-effects regression only two were longitudinal designs (Cantin et al., 2014; Angell et al., 2018), while the remainder had other nested data structures (e.g., Dixon et al., 2019). Thus, we estimate about 35% of all articles in these journals used ANOVA-based methods for managing repeated observations, whereas less than 1% used mixed-effects regression-based methods.

Contrast of Mixed Effects Regression and Repeated-Measures Analysis of Variance

Both of these methods, mixed-effects regression and repeated measures ANOVA, are valid approaches to the analysis of repeated observations. However, they each have unique strengths and limitations that make them more appropriate in different situations. The problems that arise in evaluating change vary with data structure and trade-offs one may face in selecting from available

methods. Garcia and Marder (2017) discuss three problems common to longitudinal data: (1) correlations within the data, (2) irregularly timed measurements, and (3) missing data. Correlations within data can exist when individuals are measured repeatedly over time or when individuals are clustered within the data. Irregular timing can result from variations in measurement of planned time points, such as actual measurements of a 1-month follow-up ranging from three to seven weeks, and from designation of critical events such as hospital discharge, achieving a motor milestone (e.g., walking), or the end of a competitive season. Finally, missing data are a reality for virtually every study, with implications that vary by type of missingness, data structure, and analytic method.

In weighing the effects of these different constraints, Garcia & Marder (2017) discuss other methods beyond mixed-effect regressions and repeated measures ANOVA (i.e., change scores, multivariate ANOVA, and generalized estimating equations), but the key distinction between ANOVA and other methods is that ANOVA addresses questions of *differences*, but does not explicitly model *time*. In contrast, mixed-effects regression explicitly models *trajectories* that are fit to the available data. Modelling these trajectories also means that mixed-effects regression can account for the correlations between data points within a person (as data tend to be more correlated for more proximal time points). Similarly, by explicitly modeling time as a continuous variable, mixed-effects regression can account for timing irregularities that are inherent in most research. While RM ANOVA treats repeated measurements in an arbitrary way (e.g., Time 1 versus Time 2), linear mixed-effects regression explicitly models important variability in time (e.g., the second time point may be Day 15 for one participant and Day 21 for another). Missing data is an issue for all methods, but the methods differ in how they are affected by, or capable of, addressing the consequences of missing data. In linear mixed-effects regression, missing data can reduce statistical power and produce bias. Data that are *missing completely at random* reduce power but do not bias parameter estimates. Data that are *missing at random* is when the outcome is subject to missingness, but this missingness can be accounted for by variables included

in the model (e.g., younger participants tend to be missing more data). In contrast, data are *missing not at random* when outcome values are missing due to the values in and of themselves (e.g., depression measures are more likely to be missing for more depressed individuals; Curran et al., 2010). In contrast to ANOVA, mixed-effects regression can handle missing data but still relies on the data satisfying the missing at random assumption (for missing not at random approaches, see Enders, 2011).

A summary of the conceptual differences between repeated measures ANOVA and mixed-effects regression for longitudinal data is presented in Table 1. In general, we argue that mixed-effects regression is a preferable approach for modeling change over time for its advantages in: (1) correlations within the data, (2) irregularly timed measurements, and (3) missing data. However, employing mixed-effects regression is not suitable for all situations and does come with trade-offs (Garcia & Marder, 2017; Molenberghs & Verbeke, 2001). To use mixed-effects regression, researchers need additional knowledge of how to specify appropriate models and need to appreciate the implications of choices made in the model-building process. Researchers also need access to adequate computing resources and the skill to use them. Fortunately, resources are readily available to develop the requisite knowledge and skill (Long, 2012; Singer & Willet, 2003).

Table 1. Comparison of strengths and limitations of repeated measures analysis of variance and mixed- effects regression for modeling longitudinal data.

	RM ANOVA	Mixed-Effects Regression
Model concept	• Compares means of a continuous outcome stratified by one or more categorical variable(s) to the grand mean. • Individuals are treated as a factor with error aggregated from each individual's mean of repeated measures and partitioned as intra-individual variance from the error term.	• Accounts for correlations in data with clustered or nested structure. • In longitudinal models, change over time is considered a within-person factor accounting for within-person correlations across time points and estimating error as residuals from each individual's trajectory, and between-person error is accounted for as random effects in a correlation matrix, which can be explained by fixed covariate associations with trajectory parameters.
Modeling of the outcome over time	• Addresses questions about mean difference. • Time is not inherently captured in the repeated measure, instead discrete time points are treated as levels of a categorical variable with a mean for each time point. • Mean differences between time points do not represent change over time, since time is an not explicit part of the model.	• Time is modeled explicitly for the outcome variable as a trajectory of change. • The model assumes a common pattern of change for the group (fixed effects), but individuals can vary from that pattern (random effects). • The shape of the trajectory is determined by fitting progressively more complex mathematical functions that are likely to fit the pattern of raw data scores, and testing a fit statistic (e.g., Akaike Information Criterion or Bayesian Information Criterion). • Of particular use is the ability to estimate the magnitude and timing of a plateau or other milestone on the trajectory.
Variability in timing of data points	• Requires common, discreet time points; variability in actual timing may contribute to measurement error in categorized time points. • Measurement error may accrue within time points if outcome measurement varies by time within a time point, e.g., measurement at a time point varies by ± time units around that point. Individuals' scores on an increasing trajectory may be overestimated if captured before the time point or underestimated if captured after.	• Can accommodate variability in spacing of time points and in the actual timing of individual data collection. • Time points can be spaced farther apart where little change is expected, and closer together where more change is expected • Individual measurement can vary from the target time points. If, for example, 5 weekly measurements are planned over 4 weeks, a time variable defined in days can capture the actual day of measurement, rather than collapsing to the weekly time point.

Data missing on the outcome	• Missing outcome data cannot be accommodated, without complicated statistical adjustments (such as multiple imputation) when data are missing at random. • Including only cases with complete data will reduce statistical power and risk bias to the model if data are missing not at random (MNAR). • Depending on the method employed, imputing missing values may not bias parameter estimates, but may reduce standard errors risking Type I errors in hypothesis tests.	• Data that is if missing at random (MAR) can be accommodated without excluding cases. • However, models can be biased if important time points are missing (e.g., no data where important change occurs). • Models with data that is MNAR can be fit, but models may be biased. For example, an unbalanced data set is one in which later time points are more likely to be missing, which can occur due to drop out, or outcome measurement that is performed during an intervention that varies for individuals. • Imputation of outcome data is not recommended.
Data missing on covariates	• Missing between-person covariate data cannot be accommodated. • Cases are either dropped from analysis or retained by imputing missing values.	• Missing between-person covariate data cannot be accommodated. • Cases are either dropped from analysis or retained by imputing missing values.
Time-varying covariates	• Time varying covariates cannot be accommodated in a RM ANOVA model.	• Time varying covariates can be included, but you need to careful about collinearity and variance at both the between- and within-subject levels.

Our goal in this commentary is to illustrate the advantages of mixed-effects regression over repeated measures ANOVA for researchers studying questions of change over time in motor learning and development. This article is meant to be an introduction to the topic and more thorough treatments are available elsewhere (Long, 2012; Raudenbush & Bryk, 2002; Singer & Willet, 2003). Some aspects we will discuss in detail such as the interpretation of fixed- and random-effects, but others (such as model comparison, methods of estimation, and truly nonlinear models) are topics that we will only address superficially in the interests of space. We hope that by providing an illustrative example we will motivate researchers to consider using mixed-effects regression, when appropriate, and to acquire the skills to do so in their own studies. Furthermore, by providing reproducible data and code for implementing these models in the open source software environment R (R Core Team, 2019; Bates, Maechler, Bolker, & Walker, 2015; Wickham, 2016), we hope to give researchers an affordable approach to start using these models. For users less familiar with R, however, it is important to point out that mixed-effects regression models can be implemented in most major statistical software packages. All data and code necessary to recreate the analyses below are available from: https://github.com/keithlohse/LMER_v_RM_ANOVA.

Simulated Dataset

In order to create a data set that presents some of the structure, patterns, and problems likely to be encountered by motor learning and development researchers, we chose to emulate data from a retrospective study of developmental trajectories in elite skiers (Cowan et al., 2019; Fawver et al., in press). These simulated data emulate the real data set with respect to fixed-effects and variance components, but we chose to simulate a comparable data set (N = 170 participants, k = 830 observations collected at yearly intervals) so that the data could be freely shared and disseminated. The outcome variable is United States Ski Association (USSA) points, which are used nationally to rank competitors, establish start orders, and score races, with lower scores indicating better performance. As

shown in the spaghetti plot in Figure 1A, skiers' USSA points generally reduced (i.e., improved) over time, but the rate of reduction depended on the age of the participant in 2018. We centered the Time variable on 2018 to represent the most current rankings. Visually, older participants tended to have lower intercepts in 2018 and flatter slopes, suggesting a greater rate of improvement for younger athletes early in their career. Alternatively, these data can be presented more discretely, as shown in the boxplots by year in Figure 1B. Although information about individual trajectories is lost, this method of presentation can be very useful for showing measures of central tendency and spread at the group level.

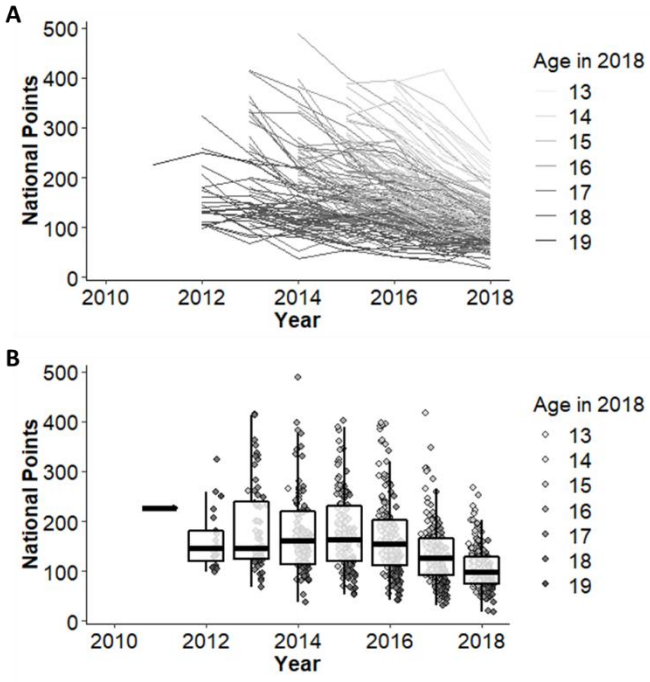

Figure 1. (A) Spaghetti plot showing the improvement over time with a line for each participant. The spaghetti plot is visually consistent with the linear mixed-effects regression approach, where a unique regression line is fit for each participant. This can be contrasted against (B) where a boxplot is shown for each year, with USSA points color-coded based on the age of each participant in 2018. The boxplots are visually more consistent with the ANOVA approach, in which group means at each timepoint are compared against the grand mean.

Statistical Analysis

In the subsequent sections we will focus on the interpretation of our models with respect to the statistical significance of their parameters and their substantive interpretation for applied researchers. However, we want to stress that evaluating a mixed-effects regression model is a complex process with steps and terms that might be unfamiliar to researchers. We refer readers to more thorough discussions of the topic that we will heavily summarize here (Long, 2012; Singer & Willett, 2003). As a useful analogue, we want to focus on the similarities between "traditional" ordinary least-squares regression and mixed-effect regression. In least squares regression, parameters are estimated to minimize the residuals. In mixed-effect regression, parameters are estimated to minimize the *deviance*. Deviance is related to the residuals in that it is a measure of error, but it is a more generalized form of the least-squares. This generalizability, however, comes at the cost of increased complexity.

One form of complexity that users are likely to encounter is the difference between *full* maximum likelihood estimation and *restricted* maximum likelihood estimation. As their names imply, both of these methods estimate the parameters that are most likely to have led to the observed data, but they differ in how they reach that conclusion. Full maximum likelihood takes all fixed- and random-effects (defined below) of the model into account, whereas restricted maximum likelihood focuses specifically on the fixed-effects (while treating the random-effects as a "nuisance" parameter through a transformation of the full likelihood function). The method a researcher uses will largely depend on their research question. Importantly, maximum likelihood estimation allows researchers to compare the relative deviance of models that have different fixed- or random-effects. These sorts of model comparisons are a common problem for applied researchers as we often want to compare models that have different fixed-effects ("What happens if I control for age?") or random-effects ("How much variation is there in individual trajectories?"). As such, all of our analyses will use full maximum likelihood estimation and we will use a model comparison approach to make decisions about parameters.

In contrast, however, an experimental researcher might know exactly which factors they want to include in their model (i.e., the experimental manipulations), so model comparison is less of a concern. In those circumstances, it would be acceptable and perhaps even preferred to use restricted maximum likelihood estimation. This is especially true in smaller samples, because it has been shown that full maximum likelihood is biased to underestimate variance components in small samples (Kenward & Roger, 1997). Thus, full maximum likelihood enables model comparisons, but is biased in small samples. Restricted maximum likelihood avoids this bias but cannot be used to compare models with different fixed-/random-effects.

Unconditional Models of Time. In mixed-effects regression, change in each individual's USSA points can be represented as a trajectory. In various texts, the equations for these trajectories are broken into multiple levels (hence the terms "multilevel" or "hierarchical" model). In our example, data points and time at "Level 1" can vary within individuals. One level higher, the slopes and intercepts at "Level 2" vary between individuals. Within a single individual, the regression equation is like that for ordinary least-squares regression, but the data are indexed by both time point (i) and by participant (j). *Residual error* (ε_{ij}) is represented as the within-person difference of each individual's observed scores (y_{ij}) relative to their trajectory:

Eq 1. $$Level\ 1: y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij}) + \varepsilon_{ij}$$

The parameters that define the average "group-level" trajectory are represented by the fixed-effects of the model (γ's). The variability of individual's parameter estimates are in turn represented by the random-effects ($U_j's$). Similarly, error at the between-person level ("Level 2") is represented by a variance-covariance matrix of the random-effects.

Eq 2. $$Level\ 2\ Intercepts: \beta_{0j} = \gamma_{00} + U_{0j}$$

$$Level\ 2\ Slopes: \beta_{1j} = \gamma_{10} + U_{1j}$$

The equations can be written together in a single equation that combines the fixed-effects and the random-effects. Fixed-effects (the γ's) are the group-level effects which can be roughly interpreted as the average intercept or the average slope across participants. In contrast, the random-effects (the U's) are the individual deviations away from the group-level effects. Finally, the residuals (the ϵ's) are the difference between the model's predictions and the actual data at each time point.

Eq 3. $$y_{ij} = \gamma_{00} + U_{0j} + \gamma_{10}(x_{\mathrm{ij}}) + U_{1j}(x_{ij}) + \epsilon_{ij}$$

This final mixed-effect equation most closely resembles the syntax that is entered into statistical programs. Although we specify only a single x-variable for time in the equation above, it is possible to add other fixed-effects to the model. For instance, we could add a curvilinear effect of time to the model (x_{ij}^2, x_{ij}^3, or even x_{ij}^4 depending on the number of data points available) or even adopt a truly nonlinear approach to the effect of time (discussed briefly below; see also Pinheiro et al., 2018; Lindstrom & Bates, 1990).

Properly accounting for the random-effects is key for drawing statistical inferences about the fixed-effects. Other conceptual definitions of fixed- and random-effects exist (see Gelman & Hill, 2007), but we think a useful definition for practitioners was advanced by Green & Tukey (1960): fixed-effects are those variables for which all levels of interest are represented; random-effects are those variables that are only a sample of larger population of potential values. For instance, in a randomized controlled trial, I would have a fixed-effect of Group (Treatment vs. Control), because those are the only two levels I am interested in, but a random-effect of Subject, because my subjects are a sample of a larger population. Depending on the research question, the same variable could be treated as either a fixed- or random-effect (e.g., treating stimulus type as a random-effect; Judd, Westfall, & Kenny, 2012).

In order to determine how best to represent time, we need to compare models with different fixed- and random-effects of time. There are different strategies for comparing models and different metrics by which they can be compared. Here, we recommend an approach of visually inspecting the

data, then starting with the linear effect and progressively adding the higher order fixed- and random-effects of time to our model (Long, 2012). In order to compare between models, we want to choose the model with the lowest deviance (i.e., the best approximation of the data). The simplest way to test the reduction in the deviance is with a chi-squared test. However, a concern as our model grows is that we might be *overfitting;* adding variables that do reduce the deviance, but at the expense of the model's simplicity and generalizability. (E.g., If I add a fixed-effect of "Subject ID Number" to my model, I will explain my data very well and with many parameters, but that model won't generalize to a new sample of subjects.) As such, we recommend that researchers use the Akaike's Information Criterion (AIC) when comparing between models (Long, 2012). The AIC introduces a penalty for additional parameters, to reduce overfitting, and thus is a more conservative approach than the chi-square test alone (Vrieze, 2012). In the simplest case, the model with the lower AIC should be selected as it is a better explanation of the data, although there are modifications to the AIC and other metrics can be used.

In our ski data example, we explored models with linear, quadratic, and cubic effects of time, see Table 2. Comparing the linear to the quadratic models, the quadratic effect of Time significantly improved the fit of the model, $\chi^2(1) = 9.36$, $p = 0.002$. The reduction in the AIC agreed, $\Delta AIC = -6.54$, suggesting that addition of the quadratic term explained more variance than it added in complexity. In contrast, adding a cubic term to the model did not reduce the deviance to a statistically significant degree, $\chi^2(1) < 0.001$, $p = 0.981$, and led to an increase in the AIC, $\Delta AIC = +2.00$. An increase in the AIC suggests that a parameter added unnecessary complexity to the model. Additionally, we tested a truly nonlinear model of time that followed a negative exponential function. This nonlinear model also led to an increase in the AIC relative to the quadratic model, $\Delta AIC = +471.99$. Note that a visual inspection of the data suggest that these data do not follow an exponential function, hence the very large AIC, but we wanted to fit this model for illustrative purposes that we will return to in the Discussion. For all subsequent analyses, we decided to retain a curvilinear model with linear and quadratic effects of time.

Table 2. Comparison between linear, quadratic, and cubic models.

Model	Fixed-Effects	Random-Effects	df	AIC	Deviance
1	Time	Time	6	8242.5	8230.5
2	Time + Time^2	Time	7	8236.0	8222.0
3	Time + Time^2 + Time^3	Time	8	8238.0	8222.0
4	Negative Exponential	Rate, Asymptote	6	8707.9	8693.9

*Note that we also tested a model with linear and quadratic random-effects of time, but that model generated a warning indicating that the fit was singular, i.e., one of the variance parameters was very close to zero on the boundary of the feasible parameter space. As such, the quadratic random-effect was dropped, leaving Model 2 as the most parsimonious model. AIC = Akaike's Information Criterion.

Conditional Models of Time. With the unconditional model for time selected, we can now add conditional fixed-effects to our model. That is, we have a very good approximation of how USSA points change overtime, regardless of other factors. Now, however, we need to see which factors explain variance in the intercepts and the slopes of our model. For instance, it is very plausible that skiers who spend more time in practice per year will have better ranks, controlling for age (Hodges et al., 2004). In our simulated dataset, we have the practice hours per year (in units of hundreds of hours) that the skiers estimated they spent in ski practice that year (as would be self-reported in a practice history questionnaire; Hendry et al., 2018). From these yearly totals, we can estimate several important values. First, we can calculate a between-subjects ("Level 2") variable which is the *grand* mean-centered number of hours per year, H_j:

Eq. 5. $$H_{between_j} = \bar{x}_j - \bar{\bar{x}}$$

This variable, which we will refer to as "*Hours.Between.c*" reflects how much a person practiced on average (in hundreds of hours per year) relative to the sample.

However, there is also variability within a person over time, so we calculated a within-subjects ("Level 1") variable, which was *group* mean-centered on each individual person's average hours:

Eq. 6. $$h_{within_{ij}} = x_{ij} - \bar{x}_j$$

This variable, which we will refer to as "*Hours.Within.c*" reflects how much a person practiced in a given year, relative to their personal average. Used in this way, hours-within is a time-varying covariate that we can use to try and explain residual variance at Level 1 of the model. Importantly, by grand mean-centering the between-subjects variable and group mean-centering the within-subject variable, we have two different measures of hours of practice that are uncorrelated, avoiding potential collinearity.

With these two variables in place, we can build a mixed-effect regression model to test our major hypotheses. The effects of Year, Year^2, Age, *Hours.Between.c*, and all of their interactions were added as fixed-effects. We also included the time-varying covariate of *Hours.Within.c* as a fixed-effect to explain residual variance in the model. Details of this model are presented in Table 3.

These effects can be interpreted as in a traditional regression output. The intercept is the estimated number of USSA points when all the x-variables are equal to zero. The time-variable of year was centered on 2018 and all other continuous variables were centered about their respective means. As such, the estimated intercept of 112.99 reflects the estimated USSA points that a skier of the average age (15.88 years) and hours of training (642.47 hrs/year) had in 2018.

The estimated slopes show the predicted change in the dependent variable for a 1-unit change in the explanatory variable. As year was centered on 2018, the negative coefficient for Year means that for every year we move backward (2017-2018 = -1), the estimated USSA points for that individual increase by -21.18*(-1) = 21.18 points.

The main-effects of Age.c and Hours.Between.c show the effects that those variables have on the intercepts. As shown by the coefficients in Table 3, there was negative relationship between Age and USSA points in 2018, with older participants tending to have better ranks (i.e., lower intercepts; Figure 2A). Similarly, there was a negative relationship between Hours.Between.c and USSA points in 2018, with individuals who had practiced more hours over their career tending to have better ranks (i.e., lower intercepts; Figure 2C).

Table 3. Results of the conditional mixed-effects model.

Random Effects

Groups	Name	Variance	Std.Dev.	Corr
Subject	Intercept	229.5	15.15	
	Year	105.9	10.29	0.23
	Residual	348.3	18.66	

N = 832 observations in k = 170 individuals.

Fixed Effects

Name	Estimate	SE	df	t-value	p-value
Intercept	112.99	2.23	338.82	50.64	<0.001
Year	-21.18	2.11	660.77	-10.06	<0.001
Hours.Between.c	-31.75	1.98	245.80	-16.01	<0.001
Age.c	-6.88	1.52	229.32	-4.52	<0.001
$Year^2$	1.41	0.49	596.77	2.88	0.004
Hours.Within.c	-18.34	1.78	560.81	-10.31	0.001
Year x Hours.Between.c	12.49	2.01	650.98	6.21	<0.001
Year x Age.c	2.44	1.32	571.56	1.85	0.065
Hours.Between.c x Age.c	4.48	1.46	234.05	3.06	0.002
$Year^2$ x Hours.Between.c	0.60	0.50	587.03	1.20	0.230
$Year^2$ x Age.c	-0.28	0.28	656.10	-0.98	0.329
Year x Hours.Between.c x Age.c	-0.97	1.29	586.37	-0.75	0.451
$Year^2$ x Hours.Between.c x Age.c	-0.28	0.31	662.88	-0.91	0.365

Note that Age was centered around the mean age in 2018, such that higher values indicate relatively older individuals. Year and $Year^2$ were also centered on 2018, so that intercepts in the model correspond the estimated rank in 2018. Estimation of all fixed- and random-effects used Maximum Likelihood estimation.

Interactions between Year x Age.c and Year x Hours.Between.c show the effects that these variables have on individual slopes. As shown by the coefficients in Table 3, there was a positive relationship between Age and the effect of Year, such that older participants tended to show flatter (i.e., less negative slopes) than younger participants, controlling for their other variables. This result suggests that younger skiers generally improved at a faster rate than older skiers. Similarly, there was a positive relationship between Hours.Between.c and the effect of Year, such that participants who spent more time in practice per year tended to have flatter slopes than skiers who spent less time in practice,

controlling for the other factors in the model. This finding might seem counter-intuitive, but it is important to remember that (A) these data were simulated, so we don't want to interpret them too much anyway and (B) the effects on the slopes need to be interpreted in light of the effects on the intercepts. More hours of practice per year were associated with less improvement from year to year, but they were also associated with higher ranks in 2018, controlling for the other variables. As such, this pattern of results suggests diminishing returns in the relationship between improvement and time in practice. An additional 100 hours of practice confers a larger benefit to someone of lower rank, but a higher ranked individual (who tends to practice more) will see much less of a benefit for an additional hundred hours of practice. This finding fits with the reality that increased effort is required to make progressively smaller gains as one rises to higher levels of competition.

Finally, we also need to interpret the effect of the time-varying covariate Hours.Within.c. This predictor explains residual variance at Level 1 of the model. That is, after accounting for the variation in the individual slopes and intercepts, there is still error in our model's prediction at any given time-point. A time-varying covariate allows us to see if any of this residual variation can be explained. In our case, there is a negative relationship between hours practiced in a given year and USSA points in that year. Controlling for the other factors in our model, for every 100 hours of practice above a person's average in a given year, our model estimates that their USSA point-ranking will improve by about -18.34 points.

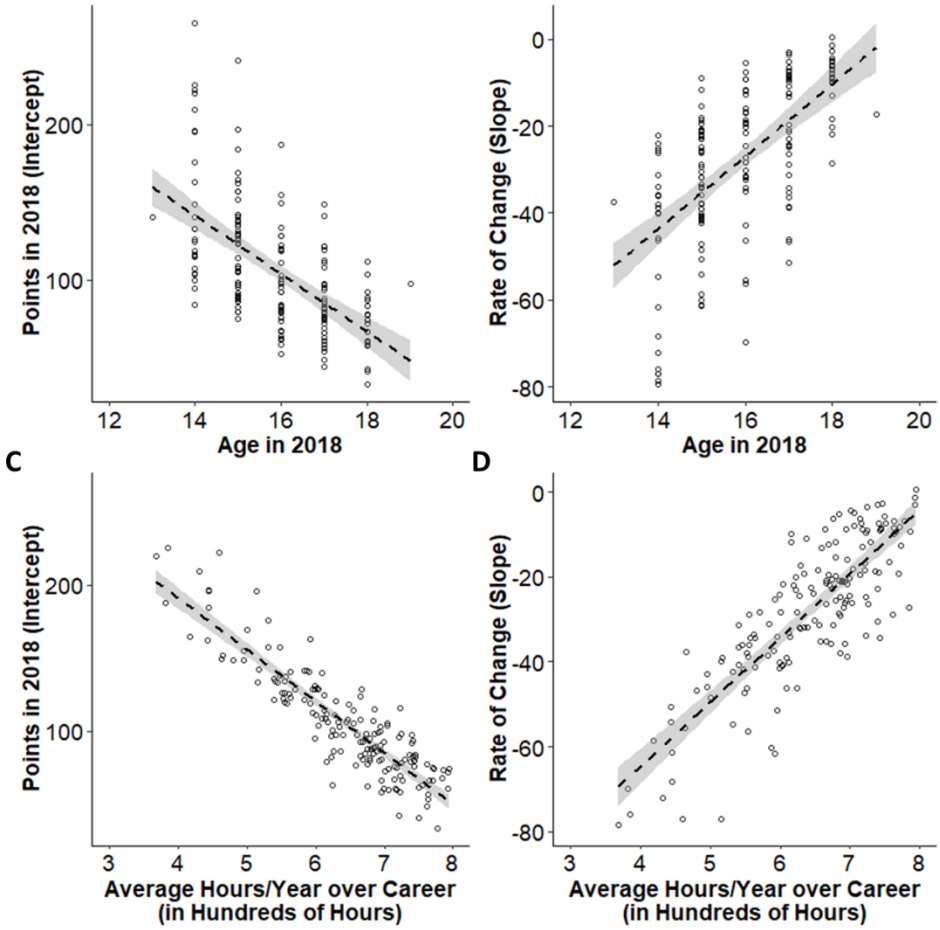

Figure 2. Individual intercepts (rank in 2018) and slopes (rate of change) as a function of (A/B) age in 2018 and average hours of practice per year over each person's career (C/D).

From this abbreviated introduction, we think mixed-effects regression provides many benefits as a method for studying longitudinal change. Mixed-effects regression allows researchers to explicitly model variation over time (at Level 1), variation between people (at Level 2), and cross-level interactions to see how the characteristics of different people affect their trajectories. Furthermore, characteristics of these data not only make the mixed-effects regression model useful but render the repeated measures ANOVA model difficult if not impossible. For instance, differences in the variability of the dependent variable over time would mean that a correction for heterogeneity of variance would need to be implemented (e.g., Greenhouse & Geisser, 1959). Different numbers of assessments per individual

require that individuals, observations, or both be trimmed from the dataset to create “complete” data. Finally, if a researcher wanted to measure the influence of a time-varying covariate, this is not implementable in repeated measures ANOVA at all.

Comparison of ANOVA to Linear Mixed-Effects Regression in a Restricted Dataset

Due to the different number of observations per person, repeated measures ANOVA was not a viable analysis method for our full data set. In order to directly compare ANOVA to a linear mixed-effects regression model, we need to restrict our dataset to a “complete” dataset with the same number of observations per person, regardless of their age. Naturally, researchers should be cautious with this kind of truncation to expediently deal with missing cases, but for the purposes of our commentary truncating the data is informative for directly comparing the two analysis methods. Excluding individuals without four years of observations and excluding cases of individuals with >4 years of observations produced a data set with 144 individuals and 576 observations. The truncated data are shown in Figure 3. Note that in these truncated data, the quadratic effect of time no longer improved the fit of the model, so only a linear effect of Year was retained as both a fixed- and random-effect.

Another key difference between these analytic methods is that mixed-effects regression allows a researcher to look at explanatory variables either continuously or as categorical factors, whereas repeated measures ANOVA requires that explanatory variables be categorical factors. To accommodate this need, we converted Year into a factor with four levels (2015, 2016, 2017, 2018), and using median splits we converted Age (Younger, Older) and Hours.Between (Low, High) into categorical factors with two levels. Dichotomizing variables using median-splits is not recommended due to the negative effects on statistical power (McClelland et al, 2015), but as with the truncation of our dataset, it is illustrative for our comparison because the numerator degrees of freedom is 1 for both a single continuous predictor and a dichotomous categorical factor. Finally, the predictor of Hours.Within.c was removed from both models as a time varying covariate is not implementable in the ANOVA.

As shown in Tables 4 and 5, we can compare the results of the two different analytic approaches with fixed-effects for the following variables: Year, Age, Hours, and all of their interactions.

Table 4. Results of the conditional mixed-effects model analogous to the repeated measures ANOVA.

Random Effects

Groups	Name	Variance	Std.Dev.	Corr
Subject	Intercept	105.89	10.29	
	Year	51.04	7.15	-0.16
	Residual	364.86	19.10	

N = 576 observations in k = 144 individuals.

Fixed Effects

Name	Estimate	SE	df	t-value	p-value
Intercept	102.44	1.80	144.03	56.82	<0.001
Year	-30.16	1.06	144.11	-28.55	<0.001
Age	-6.68	1.67	144.03	-4.00	<0.001
Hours	-31.27	1.76	144.03	-17.73	<0.001
Year x Age	3.14	0.98	144.11	3.21	<0.001
Year x Hours	10.25	1.03	144.11	9.92	<0.001
Age x Hours	4.51	1.78	144.03	2.54	<0.001
Year x Age x Hours	1.59	1.04	144.11	1.53	0.128

Note that continuous variables were mean-centered prior to analysis with the exception of year, which was centered on 2018.

Table 5. Results of the repeated measures ANOVA analogous to the conditional mixed-effects model.

Effect	df	F-value	ε_{GG}	p-value$_{GG}$
Intercept	1, 140	2039.23	--	<0.001
Hours	1, 140	137.11	--	<0.001
Age	1, 140	29.47	--	<0.001
Year	3, 420	303.55	0.84	<0.001
Hours x Age	1, 140	2.24	--	0.137
Year x Hours	3, 420	27.15	0.84	<0.001
Year x Age	3, 420	12.72	0.84	<0.001
Year x Age x Hours	3, 420	1.19	0.84	0.312

Note that Year is now a categorical factor in the RM ANOVA as opposed to a continuous variable in the linear mixed-effects regression. Within-subject effects were corrected for a violation of sphericity based on the Greenhouse-Geisser (GG) correction.

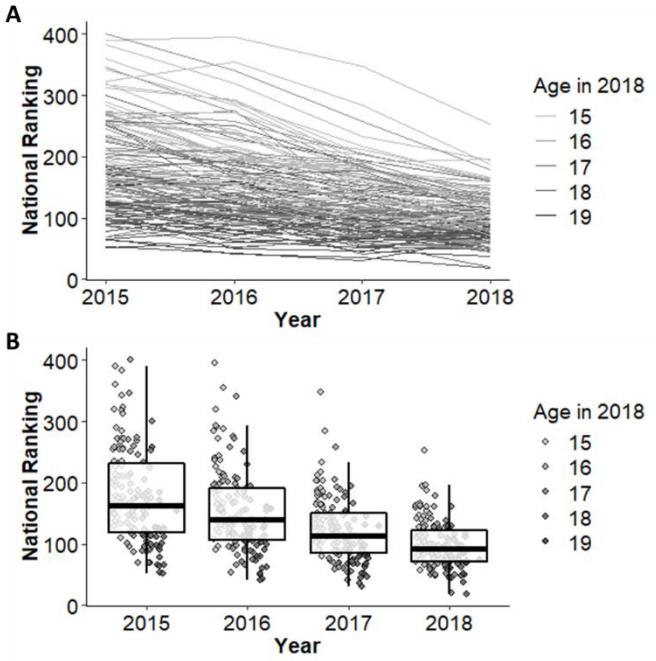

Figure 3. (A) Spaghetti plot of the truncated dataset showing the improvement over time as a continuous variable as in linear mixed-effects regression. (B) Boxplots showing mean differences where year is treated as a categorical factor, as in RM ANOVA.

Discussion

The purpose of this demonstration was to compare and contrast the strengths and limitations of mixed-effects regression and repeated measures ANOVA in modeling longitudinal data. In doing so, we have adapted a real data set of USSA point rankings for young skiers (making the dataset available to readers) and constructed three separate models for the comparison. Each model (linear mixed-effects regression with the full data set, linear mixed-effects regression with the truncated data set, and repeated measures ANOVA with the truncated dataset) highlights a different set of assumptions and trade-offs. As noted previously, the primary conceptual difference between the two approaches is that repeated measures ANOVA, compares group means against a grand mean, whereas linear mixed-effects regression explicitly models the outcome over time. Thus, neither approach is a substitute for the other.

Repeated measures ANOVA has its place in analyses where individuals may vary across measurements at multiple arbitrarily defined time points and it is important to account for within-person variance, but the omnibus test is one of group differences among those time points. In repeated measures ANOVA, 'time' is modeled implicitly only to categorize scores collected at successive arbitrarily defined points in time, as the model does not account for time intervals between the points. Conversely, mixed-effects regression explicitly models time and thus is appropriate for questions that address how individuals vary from a common trajectory over time. As parsimony is a consideration for all modeling activities, repeated measures ANOVA would be preferable where it provides a simpler solution to the question of interest.

We have demonstrated, however, that in comparison to a fully specified mixed-effects regression model, repeated measures ANOVA may be inviable, inaccurate, or provide an answer to a conceptually different question. We produced a linear mixed-effects regression model with the full data set, but repeated measures ANOVA was not a viable option due to missing data on both years and cases. Thus, we cannot compare methods on the full data set. To generate a repeated measures ANOVA, we

had to reduce the dataset using list-wise deletion. (Note, an alternative approach would be using multiple imputation to replace the missing values, but this would require a complicated statistical procedure with its own set of assumptions.) We opted to remove cases without observations for years from 2015 to 2018, and to remove all observations from years 2011 through 2014, which removed 256 observations and 26 cases. Retaining more years would have excluded progressively more cases and retaining more cases would have dramatically shortened our time frame. Additionally, if we want to include time varying covariates in the model, this necessitates the use of a mixed-effects model.

A major benefit of mixed-effects regression is that it accounts for correlations within the data and irregular measurement timing. However, improperly specified models will still be subject to distortion and bias, and there are at least two characteristics of these data that warrant further scrutiny. First, we produced two linear mixed-effects regression models, one each from the complete and truncated data sets; but which one is right? The full model includes all available data, but the dataset is unbalanced, whereas the truncated dataset is balanced, but has arbitrarily excluded cases and observations. The unbalanced nature of the full dataset may meet the missing at random assumption, in that missingness is due to age (younger skiers are not ranked in earlier years), and age is included in the model (Curran et al., 2010). However, missingness in unbalanced datasets is often associated with the values of the missing outcome variable itself. For example, models of recovery for inpatient rehabilitation samples may be missing later time points because earlier rehabilitation discharge is associated with faster recovery, or recovery to a target outcome, such as independence with mobility (Hart et al., 2014; Kozlowski et al., 2013). As such, so we cannot be completely sure that our data are missing at random even after accounting for age.

The absence of younger skiers for earlier years also hints at a second characteristic to scrutinize. Another assumption of linear mixed-effects regression is that the outcome variable is continuous. The “USSA points” that are used to rank skiers are based on the summation of “race points” and “penalty

points" for a skier's two best races each year. Race points, RP, for a single race are determined based on the following formula:

Eq 7. $$RP = (\frac{T_R}{T_W} - 1) \times F$$

Where T_R is the racer's time in seconds, T_W is the winning racers time in seconds, and F is a constant, different for each skiing discipline, based on the average spread of race results in that discipline. However, these race points are balanced against the penalty points, which consider factors like the relative ranking of the other racers in the field. For instance, the winner in a race of poorly ranked skiers will get 0 race points, but have a larger number of penalty points added to their score, whereas the winner in a race of elite skiers will get 0 race points and then have few penalty points added to their score. The calculation of penalty points is beyond the scope of our discussion (see the US Ski and Snowboard 2019 Alpine Competition Guide), suffice it to say that these points are not truly continuous, and it is harder to improve one's rank the fewer points one has (e.g., the difference between 102 and 101 is not the same as 2 and 1). In this instance, having an ordinal outcome does not have substantial consequences on the model given that our multi-level ordinal outcome has greater than 100 levels and we have a relatively large sample size (our full and truncated datasets include 170 and 144 cases, and 832 and 576 observations, respectively). However, we want to point out that if our outcomes had only a small, fixed number of levels, we could run a *generalized* linear mixed-effect model as an ordinal logistic regression. There is not, however, an equivalent alternative in ANOVA when an ordinal outcome has only a small number of levels.

Hopefully this discussion shows that the strengths of employing mixed-effects regression to model time-series data are balanced against some limitations and trade-offs. Advantages include increased control and statistical power for modeling the time parameter, options to address missingness, and interpretability of the pattern of change. Regarding interpretability, it is often easier to comprehend a curvilinear or truly nonlinear model than it is to comprehend a transformation of the

outcome or a transformation of the time variable. In our ski example, we found a curvilinear model to be superior to a truly nonlinear, negative exponential model. However, as shown in Figure 4, there are many applications in motor learning and development where researchers might find a benefit of nonlinear models.

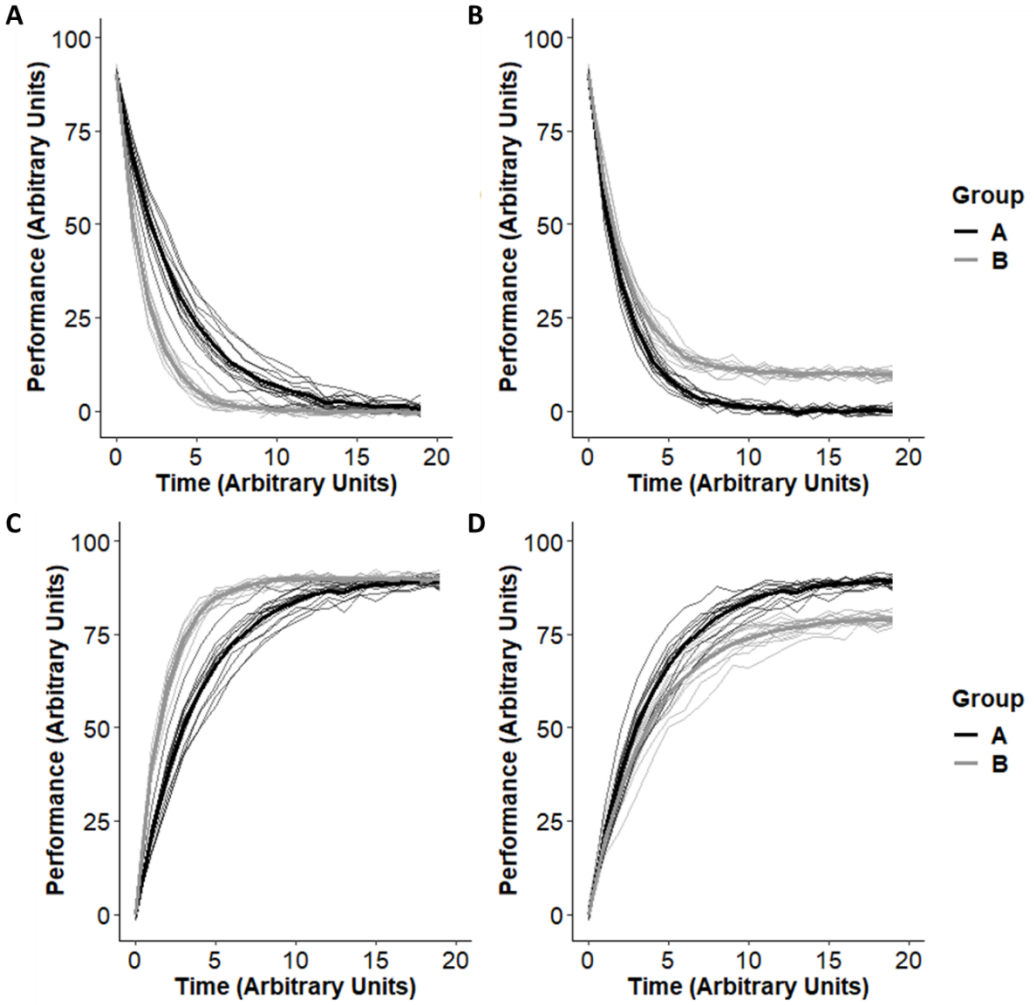

Figure 4. Simulated results illustrating how data might be modeled using a negative exponential function in a nonlinear mixed-effects model. In Panels A and B, we focus on dependent variables were negative changes in the outcome indicate improvement (e.g., error or movement speed). In Panels C and D, we show the analogous case but for dependent variables where positive changes indicate improvement (e.g., points or accuracy measures). The nonlinear mixed-effects model allows users to test for differences in both the rate of acquisition (shown in A and C) and in the performance asymptote (shown in B and D).

To illustrate this point, we simulated four different cases where nonlinear methods might be optimal due to floor or ceiling effects in the data. In Figure 4A/B we illustrate simulated data from two groups who show exponential decay as might be expected in situations where error or movement time is the outcome (i.e., there is a floor effect). In panel 4A, we can statistically model how these two groups differ in their rate of improvement. In panel 4B, we can statistically model how these two groups differ in their asymptote.

Similarly, in Figure 4C/D we can test a nonlinear mixed-effect model in a situation where our outcome might be percent accuracy (i.e., there is a ceiling effect). In panel 4C, we can statistically model how these groups differ in their rate of improvement. In panel 4D, we can statistically model how these groups differ in their asymptotes.

Naturally the nonlinear mixed model can address differences in rate and asymptote at the same time, but for the sake of clarity we have manipulated these parameters in separate samples. (The code for statistically testing these differences is provided at https://github.com/keithlohse/LMER_v_RM_ANOVA.) This sort of nonlinear relationship is common in many areas of motor learning and development, and as such nonlinear mixed-effect models have a lot of potential for researchers. Although nonlinear models are more complex than linear models (e.g., the researcher often has to provide "starting values" for the parameters to ensure the model converges on a solution; Pinheiro & Bates, 2006), they have many desirable properties such as a closer correspondence to the data and greater interpretability. For instance, we would argue it is usually easier to explain effects on asymptotes and rates of change than is to explain the effect of log-transformed 'X' on log-transformed 'Y'. Such interpretation may be important when translating results to clients, athletes, clinicians, patients, or other stakeholders.

Despite these advantages, mixed-effect models are likely unfamiliar to many readers and some of the topics we have introduced here might feel quite unusual and complex. Although resources are

available for guidance, fewer academics and clinicians have the training, experience, and tools available to employ mixed-effects regression. Although mixed-effects regression is available in most statistical software packages, the variety of mathematical functions that can be fit may be more limited. In terms of study design, fewer tools are available to estimate statistical power in mixed-effect models, but power calculators do exist (Westfall, Judd, & Kenny, 2014; Brysbaert & Stevens, 2018). Similarly, there is less consensus on how to calculate standardized effect-sizes in mixed-effect models. One can certainly calculate a proportional reduction in deviance that is like an r-squared (Singer & Willet, 2003), but the calculation and interpretation of these effects depends heavily on what random-effects are included in the model. For this reason, we would generally encourage researchers to focus on "raw" effect-sizes (e.g., β coefficients) and the precision with which they are estimated (e.g., 95% confidence intervals). Finally, it is also important to remember that determining the "best-fitting" model is based on the best-fit to the available data, not necessarily the underlying construct or theory. Inaccurate results are likely if the data do not provide a complete picture of change over time, or the available functions are not sufficiently analogous to the underlying theory. For instance, in the absence of strong theory about trajectories, a quadratic curvilinear model may be enough, but if theory strongly predicts exponential relationship, then exponential models should be tested.

Fortunately, mixed-effects regression is being applied more frequently across more academic domains. Resources are more readily available to help researchers develop the knowledge and skill to employ mixed-effects regression methodology to their specific circumstances. The capability is in many available software packages, although it may be more accessible in some than in others. In R, for example, linear and curvilinear models are readily available in the 'lme4' package, and guidance is available in a text resources (e.g., Long, 2012; Mirman, 2014). However, one will first need working knowledge and familiarity with the R programming language. Nonlinear models can also be fit with the 'nlme' package in R (Pinheiro & Bates, 2006); however, tutorials and specifications for nonlinear

functions are less readily available. Thankfully, professional development opportunities, such as online videos and webinars, introductory courses and workshops, are also becoming more common. These workshops are a valuable complement to text references on various aspects of mixed-effects regression (Long, 2012; Mirman, 2014; Singer & Willet, 2003).

Ultimately, we argue that mixed-effects regression is a valuable tool that deserves to be adopted by researchers in motor learning and development. Repeated measures ANOVA certainly has its place, but researchers need to understand what the limitations of these approaches are, and we encourage them to expand their toolkit. Mixed-effects regression provides a flexible and powerful method for understanding how individuals change over time and for modeling how these developmental trajectories are shaped by the characteristics of the person and their environment.

References

Angell, R. M., Butterfield, S. A., Tu, S., Loovis, E. M., Mason, C. A., & Nightingale, C. J. (2018). Children's Throwing and Striking: A Longitudinal Study. *Journal of Motor Learning and Development*, *6*(2), 315-332.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1).

Cantin, N., Ryan, J., & Polatajko, H. J. (2014). Impact of task difficulty and motor ability on visual-motor task performance of children with and without developmental coordination disorder. *Human movement science*, *34*, 217-232.

Cowan, R.L., DeCouto, B., Fawver, B., Lohse, K.R., Ford, R., & Williams, A.M. (2019). Developmental pathways to expertise in alpine skiers. Published in the proceedings of the *North American Society for the Psychology of Sport and Physical Activity*.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of cognition and development*, 11(2), 121-136.

Dixon, P. C., Smith, T., Taylor, M. J. D., Jacobs, J. V., Dennerlein, J. T., & Schiffman, J. M. (2019). Effect of walking surface, late-cueing, physiological characteristics of aging, and gait parameters on turn style preference in healthy, older adults. *Human movement science*, *66*, 504-510.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological methods*, 16(1), 1-16.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fawver, B., Cowan, R.L., DeCouto, B., Lohse, K.R., Podlog, L., & Williams, A.M. (in press). Psychological characteristics, sport engagement, and performance in alpine skiers. *Psychology of Sport and Exercise*.

Garcia TP, Marder K, (2017). Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. *Curr Neurol Neurosci Rep*, 17(2):14.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95-112.

Hart T, Kozlowski AJ, Whyte J, Poulsen I, Kristensen K, Nordenbo A, et al. Functional recovery after severe traumatic brain injury: an individual growth curve approach. Archives of physical medicine and rehabilitation. 2014;95(11):2103-10.

Hendry, D. T., Williams, A. M., & Hodges, N. J. (2018). Coach ratings of skills and their relations to practice, play and successful transitions from youth-elite to adult-professional status in soccer. *Journal of sports sciences*, *36*(17), 2009-2017.

Hodges, N. J., Kerr, T., Starkes, J. L., Weir, P. L., & Nananidou, A. (2004). Predicting performance times from deliberate practice hours for triathletes and swimmers: What, when, and where is practice important? *Journal of Experimental Psychology: Applied*, *10*(4), 219.

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.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.

Kozlowski, A. J., Pretz, C. R., Dams-O'Connor, K., Kreider, S., & Whiteneck, G. (2013). An introduction to applying individual growth curve models to evaluate change in rehabilitation: A National

Institute on Disability and Rehabilitation Research Traumatic Brain Injury Model Systems report. *Archives of physical medicine and rehabilitation*, 94(3), 589-596.

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 673-687.

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Sage.

McClelland, G. H., Lynch Jr, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*(4), 679-689.

Mirman, D. (2014). Growth curve analysis and visualization using R (pp. 109-112). Boca Raton, FL: CRC Press.

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, *1*(4), 235-269.

Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

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

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.