Journal of the
Experimental Analysis of Behavior

Applying mixed-effects modeling to single-subject designs: An introduction

William B. DeHart and Brent A. Kaplan

FRALIN BIOMEDICAL RESEARCH INSTITUTE AT VTC

Behavior analysis and statistical inference have shared a conflicted relationship for over fifty years. However, a significant portion of this conflict is directed toward statistical tests (e.g., *t*-tests, ANOVA) that aggregate group and/or temporal variability into means and standard deviations and as a result remove much of the data important to behavior analysts. Mixed-effects modeling, a more recently developed statistical test, addresses many of the limitations of more basic tests by incorporating random effects. Random effects quantify individual subject variability without eliminating it from the model, hence producing a model that can predict both group and individual behavior. We present the results of a generalized linear mixed-effects model applied to single-subject data taken from Ackerlund Brandt, Dozier, Juanico, Laudont, & Mick, 2015, in which children chose from one of three reinforcers for completing a task. Results of the mixed-effects modeling are consistent with visual analyses and importantly provide a statistical framework to predict individual behavior without requiring aggregation. We conclude by discussing the implications of these results and provide recommendations for further integration of mixed-effects models in the analyses of single-subject designs.

Key words: choice, humans, mixed-effects modeling, single-subject design, statistics

Behavior analysis has a long tradition of using visual inspection to describe the effects of behavioral manipulations (Parsonson, 1999; Parsonson & Baer, 1992). Visual inspection is especially well-suited to the single-subject experimental preparations often applied to the analysis of reversal and multiple-baseline/element designs. Such experimental methods are powerful in their abilities to demonstrate prediction and control (Perone, 1999), two pinnacle goals of science (Sidman 1960; see Killeen 2018 for a recent discussion). Additionally, visual analysis conveys differences of *practical* significance, that is, the magnitude of the effect and if the effect will make a difference in the individual's life. Inferential statistics, on the other hand, focus on *statistical* significance and although recent efforts have been made to highlight the importance of practical significance (e.g., effect sizes), such focus may come secondary to that of a small *p*-value (Head, Holman, Lanfear, Kahn, & Jennions, 2015).

Historically, a significant proportion of behavior-analytic researchers have rejected, or at the very least been cautious of, the incorporation of inferential statistics into the analysis of single-subject design data (Fisher & Lerman, 2014; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Shadish, Zuur, & Sullivan, 2014). Two principle concerns are typically cited. The first concern relates to the philosophical disagreement between inductive (behavior analysis) and hypo-deductive (inference from sample to population; i.e., inferential statistics) methods. This argument, although important, will not be addressed here (see Baron, 1999 for a discussion). The second concern, that of the practical implementation and interpretation of inferential statistics, will be the focus of this paper. We want to make clear that we are not advocating for the abandonment of visual analysis in lieu of statistical inference; however, visual analysis is not infallible (Danov & Symons, 2008; Fisch, 2001) and additional analytical tools would be valuable. We believe incorporating statistical techniques, *complementary* to visual inspection, is a fruitful endeavor and could further our understanding of behavior and improve our communication with other fields of psychology and funding agencies (Young, 2018a).

The focus of this paper specifically is to address two principle barriers of applying

William B. DeHart, Brent A. Kaplan, Addiction Recovery Research Center, Fralin Biomedical Research Institute at VTC.The authors would like to thank Shawn Gilroy and Jonathan Miller for their helpful insight and suggestions.

Address correspondence to: William Brady DeHart, Addiction Recovery Research Center, Fralin Biomedical Research Institute at VTC, Roanoke, VA 24016. E-mail: brady711@vt.edu

doi: 10.1002/jeab.507

inferential statistics to single-subject design data that might make behavior analysts apprehensive about using statistical inference: the compression of variability into mean scores and the aggregation of behavior across time into single data points (Branch, 1999). These analytic concerns typically result from the problematic applications of between-group inferential statistics to longitudinal data (i.e., single-subject designs). Even pseudo-nonparametric alternatives (e.g., Mann–Whitney U), by converting data points to ranks (Fay & Proschan, 2010), do not address the limitations of the application of between-group statistics to single-subject designs because by converting data points to ranks, information is further lost with all that remains being the rank order of subjects.

For many behavior analysts, exposure to inferential statistics may be limited by the relatively minimal statistical course requirements of an academic degree. Unfortunately, most introductory statistic courses—the ones most likely to be taken—focus on basic between-group inferential statistics such as the *t*-test and analysis of variance (ANOVA). These statistical methods are simple to implement and interpret, however they are only appropriate under limited experimental conditions. The conditions under which basic inferential statistics are appropriate are only when the experimental conditions match the assumptions of the statistical test. However, these assumptions are usually strict and data obtained from single-subject designs, specifically, violate these assumptions. We briefly describe these assumptions and discuss how data obtained from single-subject designs fail to meet them.

Assumptions of Standard Statistical Tests

Often times, the primary goal of a statistical test or model is to analyze the difference between means or estimate/predict an outcome from a given set of predictor variables. Almost inevitably, the model will not include all the information needed to predict the outcome with perfect accuracy. Differences between the model's predicted values and the "actual" observed values are termed "residuals." In many statistical tests, the goal is to minimize the amount of error that is "unexplained" (i.e., residuals). Greater amounts of error reduce the precision of estimates and result in more inaccurate predictions.

Normality of Residuals

The first assumption, normality, relies on the distribution (e.g., frequency) of a model's residuals approximating a normal distribution (i.e., bell curve; see top panel of Fig. 1). For standard tests, the distribution of the outcome variable is examined for normality as such instances will usually result in normally distributed residuals. Data obtained from single-subject designs may meet this assumption, but commonly, especially when rate of behavior is the outcome (e.g., counts of behavior), the distribution of behavior does not always approximate a normal distribution. Depending on the given experimental arrangement and behavior being measured, rates of behavior (e.g., aggression) may exhibit more of a skewed or truncated distribution (see bottom panel of Fig. 1) such as interventions that successfully reduce behaviors to zero or near-zero rates.

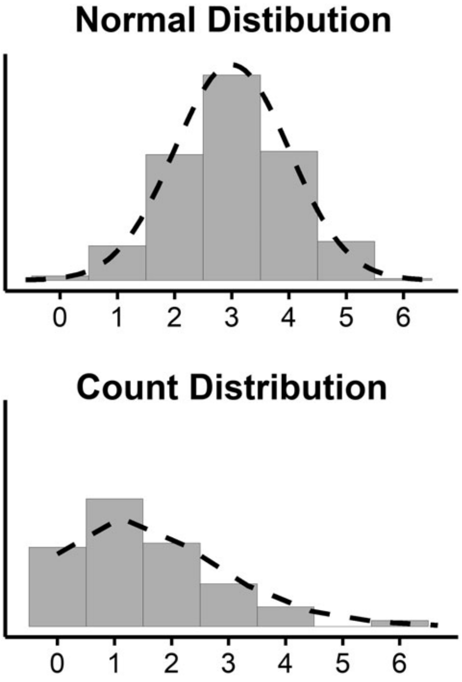

Fig. 1. Top panel: A normal distribution of scores. Bottom panel: Poisson distribution of counts.

Homogeneity of Variance

Typically, standard statistical tests rely on the assumption that the variation around the residuals' mean is similar across groups or conditions. The distribution of the test statistic (e.g., *t*, *F*) may be biased when variances, sample sizes, or exposure to different conditions are unequal. Take for example an experiment designed to decrease the occurrence of behavior. Rate of behavior may be high and variable during the baseline phase, but under a treatment phase, rates of behavior may be low. In addition, different numbers of sessions could be observed within each condition. Thus, both the variation around the means of each condition would not be similar, and sample sizes could be quite discrepant. Such an observation would violate the assumption of homogeneity of variance making typical inferential statistics inappropriate.

Independence of Observations

A final assumption that is possibly the most problematic to applications of single-subject design data is that of independence of observations. Data are independent when one observation does not influence another observation. This assumption is violated in any experimental design where the subject's behavior is sampled repeatedly. Behaviors emitted by a single subject will be more closely related compared to behaviors emitted by two different subjects. Results obtained by way of statistical tests that violate the aforementioned assumptions should be interpreted cautiously, as violations may result in biased estimates or estimates that are deemed statistically significant when, in reality, they are not (e.g., Type I error).

Mean Comparisons

Tests that compare two means, such as the *t*-test, represent the most basic aggregation of variability into means and standard errors. These tests compare two independent groups or the same group with two data points (e.g., paired *t*-test). ANOVA (or equivalently, linear regression) is relatively easy and straightforward and allows for the comparison of more than two means. With most statistical tests, the degree to which estimates will be efficient (in terms of minimizing the error) and unbiased (reflective of the "true" value) depends in part on the assumptions of the test. As discussed earlier, in a standard regression model, the following assumptions are important: 1. Normality of residuals; 2. homogeneity of variance; 3. independence of observations. However, when some of these assumptions are not met, estimates from ordinary least-squares regression[1] may not be efficient and unbiased. Indeed, because single-subject experimental designs are often set up to measure behavior over time, the standard linear regression has no way of accounting for the correlation between multiple observations within a given unit (e.g., subject), thus violating the assumption of independence.

Mixed-Effects Models

More recent developments in longitudinal statistical analyses, namely mixed-effects modeling, can aptly address the concerns of data compression both across individuals and across time (Hox, Moerbeek, & Van de Schoot, 2017). Although mixed-effects modeling (also known as hierarchical or multilevel modeling) has grown in popularity in some corners of behavior analysis (e.g., delay discounting; Friedel, DeHart, Frye, Rung, & Odum, 2016; Kirkpatrick, Marshall, Steele, & Peterson, 2018; Young, 2017, 2018b), the application to single-subject designs is limited (Baek & Ferron, 2013; Nugent, 1996). Mixed-effects models are designed to handle certain violations of assumptions in the normal regression model (Boisgontier & Cheval, 2016) that arise when analyzing single-subject design data, ultimately providing more accurate and less biased predictions of the underlying trends in the data. Like other inferential statistics, mixed-effects models provide predictive coefficients, in this case a *b* (unstandardized) or β (standardized) value, which describe the change in the dependent variable for every unit change in the independent variable. These are referred to as fixed effects and can represent both categorical and ordinal variables such as the group to which a participant is assigned (e.g., control, treatment) or a characteristic of the participant (e.g., sex, species)

[1] The most common method for linear regression seeks to minimize the model fit residuals resulting from a linear combination of predictors.

and continuous variables including time or session. Fixed effects are not unique to mixed-effects models; in normal regression models, these fixed effects are the values that are traditionally reported.

Further, interactions of fixed effects can be created to model more complex interactions such as the effects of condition or phase on behavior across time. An important characteristic of fixed effects is that the coefficient reflects an effect of a variable that is *fixed for every participant.* That a coefficient is estimated with some amount of error, typically reported as the standard error, should not be taken as a differential measure of variability for one individual compared to another; rather, the standard error is typically reflective of the variability or uncertainty in the predictor variable as related to predicting the outcome.

As a behavior analyst, one should intuitively feel that such a strong assumption —that *all* individuals in a group should be assigned a single value—is probably not appropriate. Instead, a more appropriate assumption might be that a given experimental manipulation (or treatment) will affect individuals' behavior an average of X amount, but that some individuals' behavior may be more or less responsive. Quantifying the degree to which one individual's behavior is responsive *relative to the average* is exactly the purpose of random effects. What establishes mixed-effects models as a superior alternative to simpler between-group statistics when analyzing single-subject design data is the inclusion of random effects. Random effects quantify individual subject variability without eliminating it or assigning it to an overall error term (which assumes that the individual variation around the grand mean is identical for all participants) as is done in *t*-test and ANOVA analyses (Fig. 2).

Two types of random effects can be included in the analyses of single-subject design data. First, a random intercept allows for the behavior of each individual to start at a different place (if zero on the predictor scale represents initial performance). This contrasts with the fixed-effects arrangement, which asserts that all individuals' behavior in a group or condition begins at the same level. Second, one or more random slopes allow for individual differences in the change in behavior across minute, session or condition. Again, this contrasts with the simpler fixed-effects-only models, which estimate a single slope value for all individuals' behavior in a group or condition.

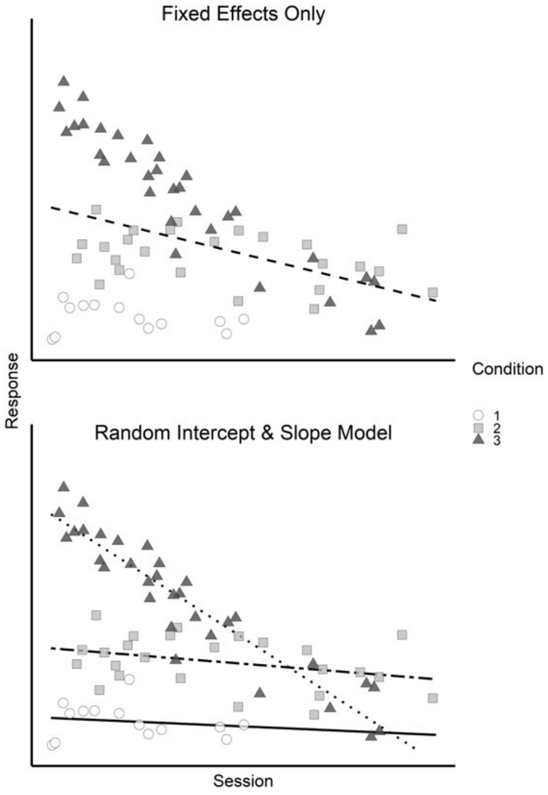

Fig. 2. Comparison of model with only fixed effects and model with random effects. Top panel: standard regression model fit to grand mean. Bottom panel: mixed-effects model with random effects for intercept (condition) and slope.

Along with fitting one regression line to the mean of all individuals (e.g., fixed effect) to predict sample or group wide behavior, including random effects, allows for the prediction of an individual's behavior to vary in the context of the larger group. For example, the fixed-effect coefficient in a hypothetical regression model may estimate that individuals in Group A display a mean difference of X amount greater compared to Group B. A random-intercept component may estimate that Subject 1 in Group A shows X amount greater compared to Group A's mean, whereas the random-intercept component may also estimate that Subject 2 in Group A shows X amount less compared to Group A's mean despite an overall difference between the two groups' means. Panel 1 of Figure 2 compares the model fit of a traditional ANOVA or regression analysis with only a fixed effect that was fitted to all participant data in Group A. Panel 2 of Figure 2 displays the inclusion of a random intercept for condition and a random slope for session. Note that the three regression lines' intercepts differ as do their slopes. Such important complexity cannot be modeled in traditional ANOVA or regression methods.

A second strength of mixed-effects modeling is the ability to account for the correlation of nonindependent data points within a defined cluster. Clustering describes data that, by group association, is related in a meaningful way. Therefore, within a data set, multiple "levels" exist. For example, when comparing classroom (Level 1) standardized exam performance within a school (Level 2), data obtained from individual classrooms are clustered together because they share the same teacher and classroom environment (Fig. 3). By analogy, data obtained from single-subject designs are also clustered. Data obtained from sessions (Level 1) are clustered together because they share the same subject (Level 2; see Fig. 3). Importantly, mixed-effects modeling is flexible in how this clustering can be accounted for. For instance, data obtained from contiguous sessions (e.g., session 1, session 2) are expected to be more correlated than data obtained from sessions farther apart (e.g., session 1, session 10). In this case, an autoregressive correlation, in which the

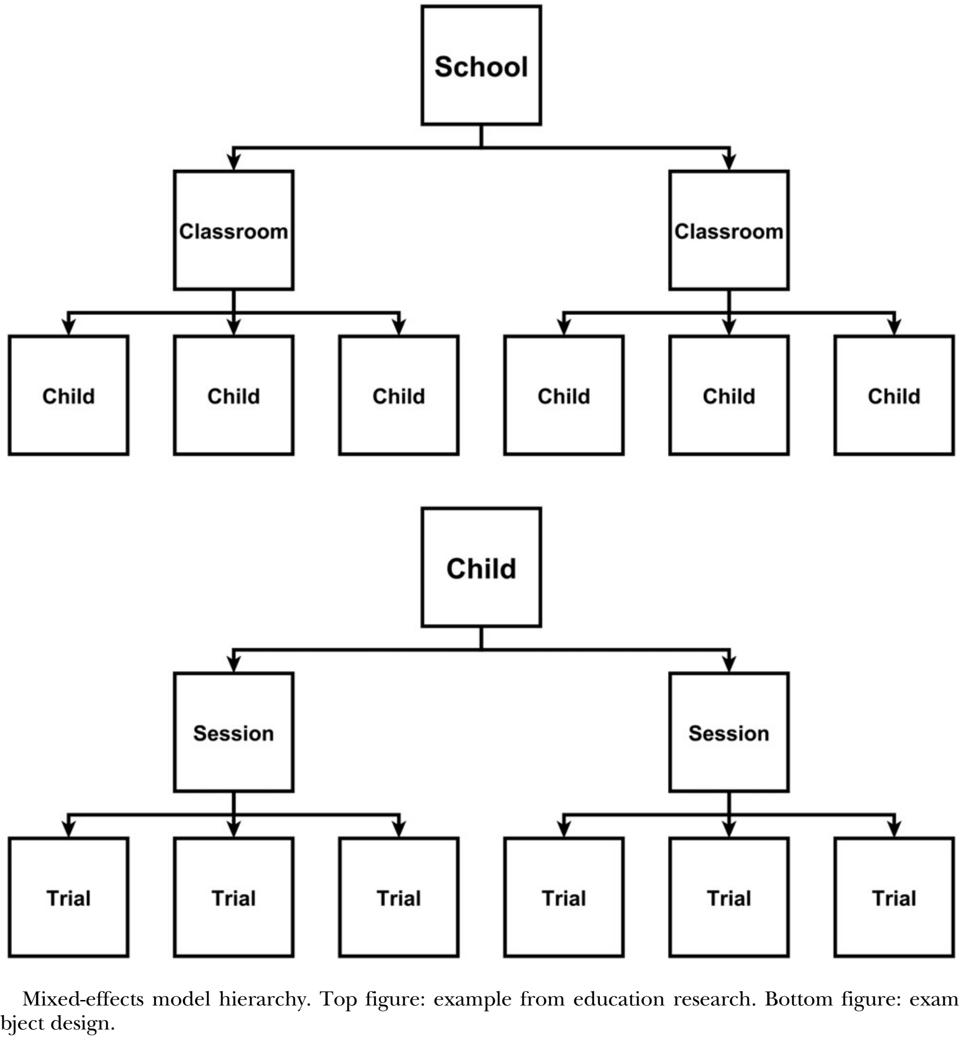

Fig. 3. Mixed-effects model hierarchy. Top figure: example from education research. Bottom figure: example from single-subject design.

correlation between data points decreases as the temporal distance between them increases, would be appropriate.

Another benefit of mixed-effects modeling is its ability to maximize the degrees of freedom by using all available data points and therefore increasing the statistical power of the test. In order to analyze data at the single subject level, one could fit a regression model to each subject's data. One could then examine the values or coefficients from the individual regression fits and conclude that subjects respond in a similar way or that a given procedure might affect subjects in a similar way (i.e., decrease behavior) but the amount of information to be used in each regression will be a fraction of the whole and could lead to large variations in estimates. In addition, by increasing the number of statistical tests (e.g., a regression model for each participant), the propensity to commit Type I error (concluding there is a statistically significant effect when there is not one) increases. Mixed-effects modeling not only estimates a "mean" change (which can be compared between groups using pairwise comparisons) but, as mentioned earlier, these models allow individual subjects' behavior to vary (e.g., random effects) from that "mean" change. Therefore, because information from all participants is used, estimates will likely be more accurate and "extreme" cases or outliers will be shrunk towards the overall mean (Kirkpatrick et al., 2018), thereby limiting the need to exclude subjects from the analyses.

Mixed-effects models can also more accurately model longitudinal changes in behavior. One assumption could be that behavior change across time occurs at a linear rate. However, change in a target behavior could also occur in a nonlinear fashion. Mixed-effects models, similar to multiple regression, can incorporate transformations of continuous predictors to model behavior that changes in a nonlinear way (Hox et al. 2017).

Finally, mixed-effects modeling allows for the analyses of nonnormally distributed data and can account for certain types of missing data using Maximum Likelihood Estimation (Krueger & Tian, 2004). This family of models is referred to as generalized linear mixed-effects models. Whereas these more complex approaches make analyses and interpretations of results potentially more complicated, model estimates are less likely to be biased and the probability of committing Type 1 error is minimized. Generalized linear mixed-effects models extend the linear mixed-effects models and can analyze count, ordinal, binomial, multinomial, zero-inflated (e.g., disproportionate number of zeros), and over-dispersed (e.g., the variance of scores is greater than predicted) data while still accounting for the information that behavior analysts find valuable (Bolker et al. 2009; Hox et al. 2017). The flexibility of generalized linear mixed-effects modeling is especially relevant to behavior analysts as single-subject designs are often measured using discrete counts of behavior, which traditional analyses such as the ANOVA are not equipped to analyze.

Although these generalized variants still model the outcome variable as a linear combination of predictors (e.g., $Y = B_0 + B_1x_1 + B_2x_2$), a link function is used to transform this linear combination onto the space of the outcome variable, especially when the outcome variable is not normally distributed (e.g., counts, rates that cannot go below 0). In addition, without a proper link function, variance of the residuals may not be constant, giving rise to biased results. Most applicable to behavior analysts are the poisson and negative binomial distributions, which address count data (i.e., whole integers; e.g., 1, 2, 3). In addition, adjustments for zero-inflation and over-dispersion would commonly be necessary to more accurately describe count data.

In order to demonstrate the power of mixed-effects modeling, we reanalyzed data from Ackerlund Brandt et al. (2015), present the mixed-model results, and indicate how they compare to the conclusions obtained through visual analysis. The purpose of this reanalysis is to provide a basic demonstration of how mixed-effects modeling can be applied to single-subject design, without needing to aggregate data into simpler units such as means and standard deviations. We also provide the access to the dataset[2] and R code (Version 3.5.1; R Core Team, 2018).

Method

Source Study Methods

To illustrate the application of mixed-effects modeling to single-subject design data, we

[2] https://github.com/brentkaplan/jeab-mixedeffects

extracted[3] data from Study 1 by Ackerlund Brandt et al. (2015). The researchers were interested in examining the degree to which choice functioned as a reinforcer. In their study, 30 typically developing children completed a choice assessment configured in a concurrent-chains arrangement. Fifteen presentations (trials) were conducted in each session. During the initial link, participants indicated whether they preferred the a) child choice (praise and five edible reinforcers presented on a plate and child chose among the options); b) experimenter choice (praise and five edible reinforcers presented on a plate and experimenter chose among the options); or c) control (praise and the presentation of an empty plate). During the terminal link, participants engaged in expressive picture labeling and if they correctly answered (unprompted or not), the outcome chosen during the initial link was delivered. The dependent variable was the frequency of selections during the initial link. For more details related to the study procedures, see Ackerlund Brandt et al. (2015).

Source Study Results

The source study did not apply any statistical tests to the data, rather the researchers utilized visual analysis to describe trends in the data. Ackerlund Brandt et al. (2015) described that two thirds (20/30) of the participants showed a preference for child choice, whereas one third (10/30) showed no difference in preference between the child and experimenter choice options. The authors described three distinct patterns among the participants who preferred the child choice: a) high and consistent preference, b) variable levels but higher overall, and c) similar preference initially transitioning into higher preference.

Data Extraction

We independently extracted the data from Figures 1 and 2 (pgs. 351-352; 15 participants total) of the source study using WebPlotDigitizer (Mani, Sharma, & Singh, 2018). WebPlotDigitizer has been shown to result in high levels of reliability and validity for extracting single case design data (Drevon, Fursa, & Malcolm, 2017). Data were first rounded to the nearest integer and 100% of the data were compared using exact interobserver agreement (Reed & Azulay, 2011). Total agreement was 100%.

Data Analysis

A two-level generalized linear mixed-effects model[4] was created to investigate the effects of choice type on choice selection in the initial link using the glmmTMB package (Brooks et al., 2017) for model fitting and the DHRMa package (Kuznetsova, Brockhoff, & Christensen, 2017) to conduct model diagnostics. By two-level, we mean that session (i.e., session 1, session 2, etc.; level 1) and condition (i.e., control, experimenter, child) are *nested* within an individual (level 2; see also Fig. 2 for an example of nesting). All analyses were conducted in the R statistical environment (Version 3.5.1; R Core Team, 2018). It should be noted that a variety of programs are capable of conducting mixed-effects modeling including SPSS ("IBM SPSS Statistics Overview," n. d.), SAS ("Analytics, Business Intelligence and Data Management," 2018), and STATA (StataCorp & Others, 2007). We chose R because it is open source and provides a variety of additional features for data preparation.

As mentioned earlier, frequency of selections served as the dependent variable. The fixed effects (independent variables) included condition (treated as a nominal factor consisting of three values; i.e., control, experimenter, child), session number (treated as an interval variable; i.e., session 1, session 2), a logarithmic adjustment for session (logsession; models the nonlinear change in frequency of selection across sessions), and an interaction term of condition by session number. For the condition variable, dummy coding was used such that two coefficients are estimated: the difference between the experimenter and control conditions and the difference between the child and control conditions. The intercept, therefore, reflects the estimated mean

[3] We were unable to obtain the original data from the corresponding author.

[4] We elected to conduct a generalized linear mixed-effects model over a linear mixed-effects model despite the increase in model complexity because the target behavior was measured in counts. Preliminary analyses indicated that the generalized model presented here provided a superior fit to a misspecified linear model.

frequency of selections of the control condition (i.e., the reference group) at session “0”.

Without the inclusion of random effects, a fixed-effects-only model would be a standard generalized linear regression. A standard model would result in an “intercept” for each condition (i.e., the mean frequency of selections at session “0”) and allow the frequency of selections to change (increase, decrease, stay the same) at different rates for the three conditions. For example, this standard model could allow the frequency of selections for the control condition to stay the same while allowing for an increase in child-choice selections. However, this model would only provide predictions for the participants as a whole group and does not accommodate individual variations around the group’s mean selections.

A random intercept of child, therefore, was specified to allow for individual variations in each child. A random slope each for condition and for session were also specified to allow for individual variation in the rate of selections over the progression of sessions and by condition. In order to determine if the random effects accounted for a meaningful proportion of variance, the interclass correlation (ICC) was calculated. The ICC is the amount of variance accounted for within a cluster (e.g., subject) by the random effect. Though no objective rule exists, researchers suggest that an ICC of at least 5% justifies the inclusion of a random effect (LeBreton & Senter, 2008). The ICC for child was 30% meaning that the random intercept accounted for a large proportion of the variance in the data. Additionally, the ICCs for the experimenter and child conditions random slopes were large at 34% and 35% respectively, indicating that the rate of change for each condition was different. However, the ICC of the random slope of log-adjusted session was less than 0.000 suggesting that the random slope for session did not account for a meaningful degree of variance in the data. Pairwise comparisons were conducted to investigate differences in the different choice conditions using the emmeans package (Lenth, 2018).

The model syntax for both a linear (lmer package; Bates, Mächler, Bolker, & Walker, 2014) and generalized linear (glmmTMB package) models can be visualized as:

choice ~ condition + log(session) + condition *log(session) + (condition + log(session) | child) where choice frequency is predicted by the interaction of condition by log adjusted session. The term (condition + log(session) | child) is the random-effect term which includes a random intercept for each child and random slopes for condition and session.

Results

Figure 4 depicts the study results at the single-subject level. A visual analysis of selections confirms that both the experimenter and child choices were chosen more frequently over the control choice. Additionally, for most participants, the child choice was selected more frequently than the experimenter choice. This interpretation is similar to that of Ackerlund Brandt et al. (2015).

The following model-fit results outline the typical steps of determining the appropriateness and quality of a model fit. First, regression diagnostics were conducted to determine the quality of the model fit (Fig. 5). The top-left panel of Figure 5 tests the assumption of linearity of the dependent variable. The regression line through the values is sufficiently straight to not violate this assumption. The top-right panel of Figure 5 displays the residuals of the model fit. For this generalized linear mixed-effects model, residuals are more appropriately investigated by analyzing the distribution of residuals at each quantile. The three dashed lines represent the pattern of residuals at each quantile and are sufficiently straight to conclude that there was no systematic bias in the distribution of residuals. The bottom-left panel depicts the Q–Q (quantile–quantile) plot, which plots the observed quantiles against the predicted quantiles. Deviations from the dotted line (slope of 1) should be minimal. Finally, the bottom-right panel depicts a histogram of the model fit residuals. The residuals are sufficiently normally distributed to not violate this assumption.

The ability of the model to account for zero inflation (i.e., a high frequency of zeros; in the current dataset, primarily driven by zero or near-zero selections of the control condition) and overdispersion (i.e., greater variability than would be expected given a certain distribution) were also assessed. The model fit was simulated (N = 1000) to create a dataset for comparisons to the actual model-fit results. In order to test the necessity of adjusting for

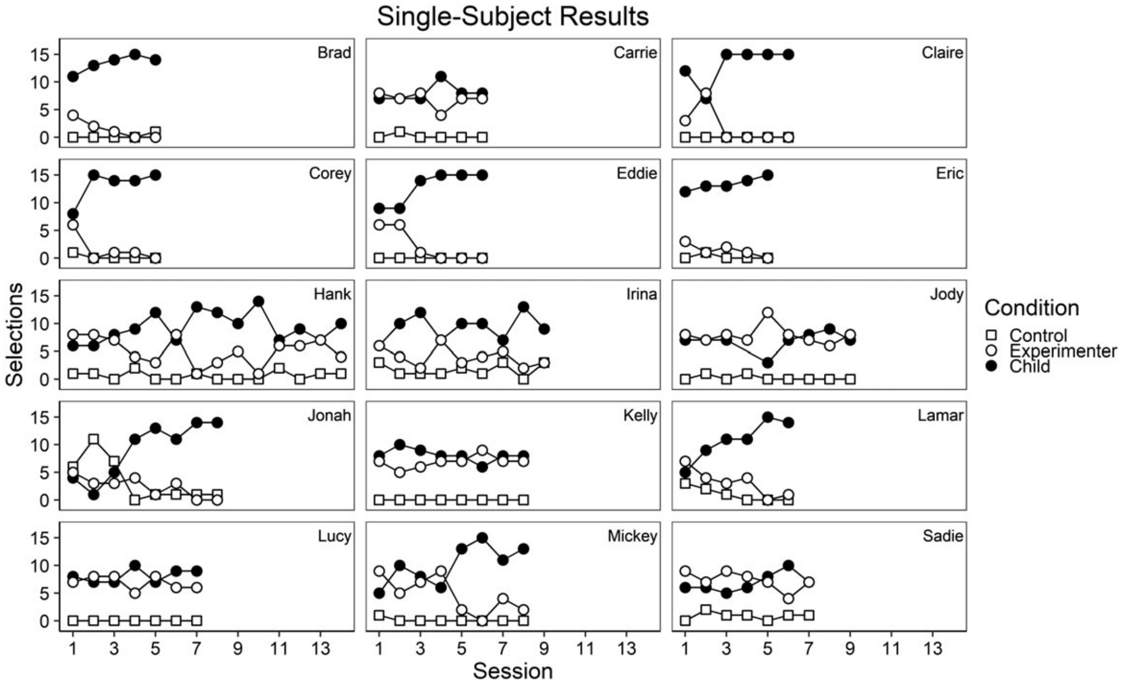

Fig. 4. Single-subject data extracted from Ackerlund Brandt et al. (2015) regraphed here to serve as reference for the mixed-effects model results and predictions.

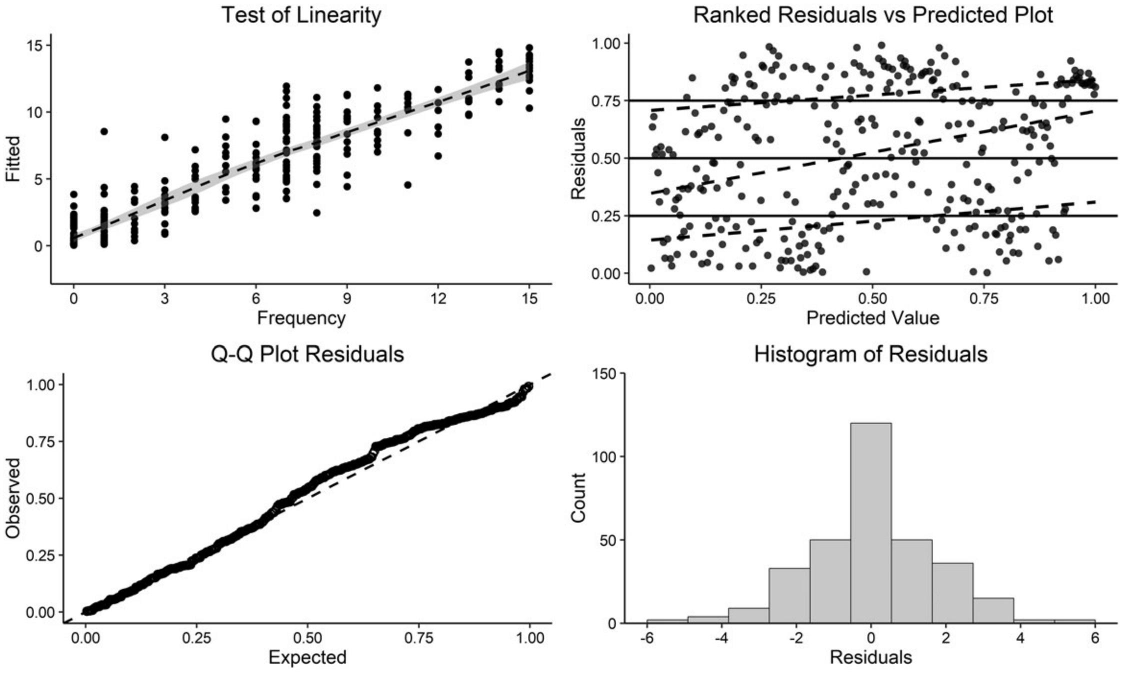

Fig. 5. Regression diagnostics. Top left: test of linearity. Top right: plot of model fit residuals. Bottom left: q–q plot of residuals. Bottom right: histogram of residuals.

overdispersion, the ratio of the outcome variance to the grand mean of the outcome was calculated with scores above 1 suggesting overdispersion (Balakrishnan et al., 2014). Overdispersion was apparent in the data (var/mean = 4.350), justifying the use of the negative binomial link function in the model, however the model successfully accounted for overdispersion in the data (overdispersion $\chi^2 = 0.923$, $p = .184$). A significant number of 0's (e.g., zero inflation) was also detected in the data (zero-inflation $\chi^2 = 460.651$, $p < .001$). The generalized linear mixed-effects model also successfully accounted for zero inflation (the predicted ratio of expected 0's from model fit to 0's from simulated model fit was 0.983, $p = .888$).

The results of the generalized linear mixed-effects model confirm the conclusions drawn from visual analysis (Table 1) and provide additional information useful to behavior analysts. The control condition served as the reference group for all statistical analyses. Regression coefficients should be interpreted as a log unit change in the expected count of the outcome. We also report the exponentially (i.e., exp) transformed regression coefficient which can be interpreted as the change in the expected outcome. Overall, the model fitted the data very well. The marginal R^2 (i.e., variance accounted for) value for only the fixed effects was 0.732 and the conditional R^2, the overall R^2 which includes both the fixed and random effects, was 0.964. A statistically significant main effect for condition was found for both the experimenter ($b = 2.457$, $p < .001$) and child choices ($b = 2.764$, $p < .001$) indicating that participants were 12 times more likely to select the experimenter choice over the control choice and 16 times more likely to select the child choice over the control choice. The main effect for session ($b = -0.639$, $p < .01$) was significant suggesting that the distribution of choices changed across time. The interaction of condition and session was also statistically significant for the experimenter condition ($b = -0.33$, $p < .001$) and the child condition ($b = 0.863$, $p < .001$). However, the direction of these effects were different, indicating that the likelihood of choosing the experimenter choice decreased across time whereas the likelihood of choosing the child choice increased across time. Finally, pairwise comparisons were conducted to analyze the overall differences between the three choice conditions. Both the experimenter ($M = 3.84$) and child ($M = 10.649$) choices were more frequently selected than the control ($M = .508$; $t(28) = 3.329$, $p < .01$; $t(28) = 10.141$, $p < .001$ respectively) choice. Also, the child choice was more frequently selected than the experimenter choice ($t(28) = 6.812$, $p < .001$).

In summary, whereas both the experimenter and child choices were chosen more frequently than the control choice initially, for many participants, the rate of selecting the experimenter choice decreased while the rate of selecting the child choice increased.

Table 1

Mixed-effects model results

Conditional Model Fixed Effects	*b*	*Std. Error*	*t-value*
Intercept	-0.758	0.496	-1.529
*log*Session	-0.641	0.175	-3.674***
Condition[Experimenter]	2.462	0.521	4.724***
Condition[Child]	2.758	0.527	5.234***
Condition[Experimenter]**log*Session	-0.332	0.189	1.760
Condition[Child]**log*Session	0.864	0.178	4.847***
Random Effects	*Variance*	*Std. Dev.*	
Child (Intercept)	2.145	1.464	
Condition[Experimenter]	2.434	1.560	
Condition[Child]	2.539	1.593	
*log*Session	0.003	0.057	

*p < .05,

**p < .01,

***p < .001. Condition[Experiment] and other predictors in brackets indicate that that predictor is being compared to the control condition.

Figure 6 displays the model predictions for each individual participant which confirms this conclusion.

Discussion

Mixed-effects modeling is increasingly popular in psychological research, and especially within behavior analysis (Young, 2017, 2018b). Although more complex than standard inferential statistics (e.g., *t*-test, ANOVA), mixed-effects modeling addresses some of the shortcomings of these traditional techniques when applied to single-subject design data. Specifically, we sought to address two barriers of applying inferential statistics, including compression of variability into mean scores and the aggregation of behavior across time into single data points. Towards this end, we applied the mixed-effects modeling technique to single-subject data collected by Ackerlund Brandt et al. (2015). The results of our analyses largely confirmed the conclusions drawn from visual inspection by the researchers and importantly, our technique did not rely on aggregating data into fewer data points. Rather, our mixed-effects model was estimated using all available data (i.e., frequency of choice selection for every child and for each condition at every session).

A greater understanding and application of inferential statistics would be advantageous to the field of behavior analysis. Other fields of psychology, insurance companies, and even parents could benefit from statistical analyses of the effectiveness of single-subject designs and behavioral analytic methods. Likewise, funding mechanisms such as the National Institutes of Health or the Institute of Education Sciences are increasingly requesting statistical outcomes to demonstrate treatment efficacy (NIH, n.d.). Recent concerns over the replicability of much of psychology has arisen as large-scale replication studies have failed to replicate many published findings (Carter & McCullough, 2014; Open Science Collaboration, 2015). While some of this can be attributed to dependence on statistical inference and "*p*-hacking" (manipulating test parameters until a desired outcome is obtained), we assert that a principal cause of the recent "replication crisis" is an over-reliance on between-group designs. Between-group designs, particularly between-group designs with small sample sizes,

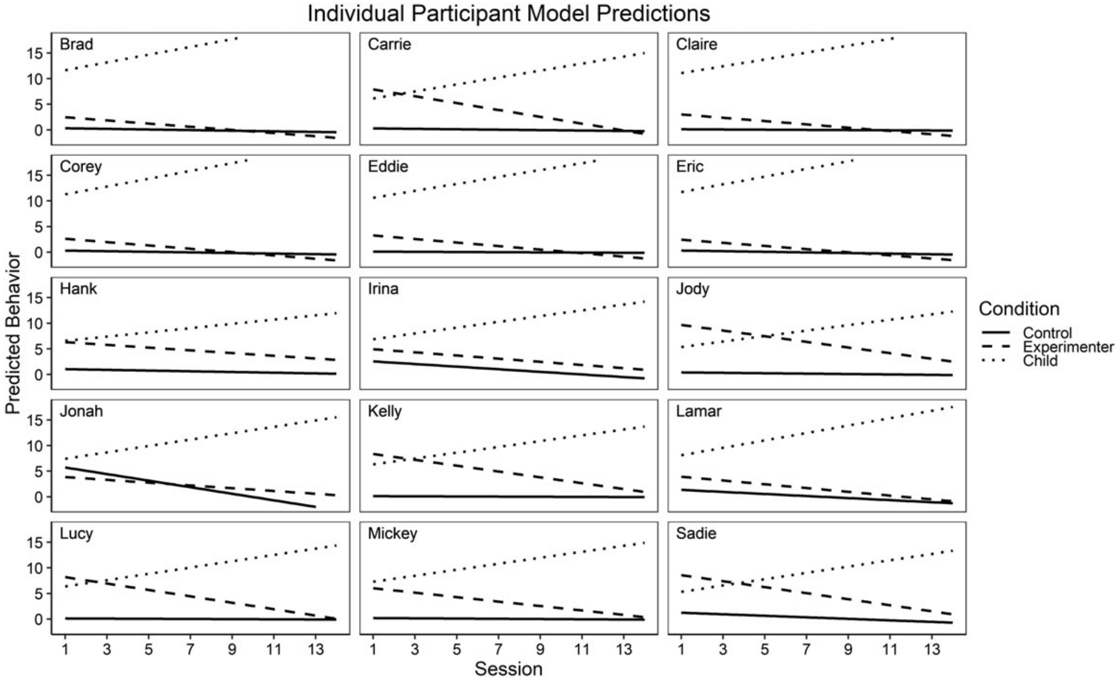

Fig. 6. Mixed-effects model predictions for each condition for each participant.

greatly increase the likelihood of committing Type I error (Banerjee, Chitnis, Jadhav, Bhawalkar, & Chaudhury, 2009).

Behavior analysis has much to offer to the rest of psychology in regard to more effective experimental design. Single-subject designs (or any within-individual variant) are powerful techniques for demonstrating experimental prediction and control (Perone, 1999). However, we contend that the influence of such methods is restricted by the inability to communicate with the rest of psychology using the common language of statistical inference. Here we present the mixed-effects model and demonstrate how it can be applied to single-subject data while addressing many of the historic issues that have limited the application of statistical inference in behavior analysis. We believe behavior analysts should view mixed-effects modeling as an *additional* tool to visual analysis.

The results of our modeling suggested that most participants preferred the child choice over the control and experimenter choices. As depicted in Figures 4 and 6, most participants demonstrated greater preference for the child choice as sessions progressed. Although examination of the original data via visual analysis suggests that some participants (e.g., Jody, Kelly) show relatively stable indifference between the child and experimenter conditions, these participants' predicted values show relatively greater changes in preference across increasing sessions. This is because the model accounts for all of the participants' data, which reflects general increases in selecting child choice and decreases in selecting experimenter choices. It may be possible that for these participants, more sessions assessed or enhanced discrimination may have resulted in greater differentiation between conditions. That is, consistent with Ackerlund Brandt et al. (2015), these results support the conclusions that the child choice condition was most frequently preferred while allowing and recognizing that individual participant behavior may differ (e.g., Jody, Kelly).

As noted earlier, these results do not replace valuable visual inspection and no science should blindly accept statistical results without a thorough evaluation of the data-collection methods and conclusions drawn from the results (Ator, 1999), as well as striving to replicate novel findings. Indeed, methods do exist to standardize and quantify the visual analysis process (see Fisher, Kelley, & Lomas, 2003) thereby reducing the likelihood of committing Type I error. We note that concordance between the visual inspection and statistical conclusions *should* be expected if the results of the mixed-effects model are to be of value to the behavior analyst, though they may not always show perfect correspondence. When they do not correspond, further investigation is necessary to determine if the issue is a shortcoming of the model or the visual inspection. However, we believe that these additional analyses complement visual analysis, increase the impact of the results, and improve our ability to communicate with the rest of psychology. Furthermore, the mixed-effects framework presented here provides better prediction (and thus allows for greater experimenter control) compared to more basic statistical tests that would otherwise be conducted (e.g., ANOVA). Although basic statistical tests would indicate how the participants (as a group) allocate their responding to the different conditions, they do not provide the flexibility to model and predict more individual deviations (from the group) in preferences such as those shown by Jody and Kelly. Taken together, the addition of the mixed-effects model analysis allows the behavior analyst to speak in the language of statistics—a set of stimuli familiar to psychologists, policy makers, and others outside of the behavior analytic field. We believe complementing traditional approaches to evaluating single-subject designs with statistical analyses "buys" the behavior analyst greater credibility with others and demonstrates the probability of obtaining a difference as large if not larger given the null hypothesis (e.g., no effect; i.e., definition of a *p*-value).

There are limitations to this specific implementation of mixed-effects modeling to single-subject design demonstrated here. First, we could only access the data visually represented by Ackerlund Brandt et al. (2015). We could not obtain the entire data set of all 30 participants. The additional participants would have improved the power of the analyses, though we do not expect the results to have changed substantially. Another limitation specific to this implementation is that this was not a multiple-baseline design. Mixed-effects analyses of designs with multiple behavior rate changes (e.g., behavior increases and decreases) or interrupted time-series designs, while possible, will be more difficult to implement. Our hope is that by encouraging the adoption of mixed-effects models for analyzing single-subject design data, greater interest in further advancing these analyses will develop.

There are also important general limitations to the implementation of mixed-effects modeling to single-subject data, which can help guide a priori experimental design preparations. Mixed-effects are more robust towards smaller sample sizes compared to ANOVA and multiple regression but they still require larger sample sizes than many behavior analysts are accustomed to use (McNeish & Stapleton, 2016). Conservative estimates suggest that a minimum of 20 clusters (e.g., subjects for single-subject designs) with approximately 20 observations (e.g., sessions) are necessary to avoid parameter biases (Austin, 2010). Bayesian implementations are also available and growing in popularity and are particularly robust for analyzing data with fewer clusters, but these analyses are more complex (see corresponding articles in this Special Issue). Nevertheless, it is not uncommon for an experiment in behavior analysis to report findings from eight or even four subjects. If researchers are to implement these techniques, larger sample sizes will be required.

Perhaps the largest limitation for many researchers in implementing these methods is the opportunity costs of mastering a complex analytic technique. While we provide the R code and sample data for the analyses, familiarity with the analyses is necessary to avoid model misspecifications. In response to this limitation, we encourage interested researchers to seek out collaborators, department statisticians, and even encourage graduate students to advance their statistical training. For those interested in a deeper understanding, we recommend *Multilevel Analyses: Techniques and Applications* (Hox et al., 2017) for a thorough introduction and *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman & Hill, 2007) for a more advanced discussion.

We believe that for many, an introduction to these methods and brief training in when to implement them and how to interpret their results would be a valuable point of growth for behavior analysis as a field. We note that although no single study will necessarily incorporate all the advantages that mixed-effects models have over ANOVA and similar analyses, we are confident that the adoption of these more advanced techniques would be valuable to the individual researcher and to the field more broadly. Mixed-effects modeling resolves many of the concerns of using inferential statistics with few compromises (e.g., slightly larger sample sizes, sometimes increased complexity), would allow for meta-analyses, and would further our influence in the rest of psychology.

References

Ackerlund Brandt, J. A., Dozier, C. L., Juanico, J. F., Laudont, C. L., & Mick, B. R. (2015). The value of choice as a reinforcer for typically developing children. *Journal of Applied Behavior Analysis, 48*(2), 344–362. https://doi.org/10.1002/jaba.199

Analytics, Business Intelligence and Data Management. (2018, August 10). Retrieved August 24, 2018, from https://www.sas.com/en_us/home.html

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *The Behavior Analyst, 22*(2), 93–97.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics, 6*, Article 16. https://doi.org/ 10.2202/1557-4679.1195

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., & Teugels, J. L. (Eds.) (2014). Overdispersion. In *Wiley StatsRef: Statistics Reference Online* (Vol. *8*, pp. 1–9). Chichester, UK: John Wiley & Sons, Ltd.

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal, 18*(2), 127–131. https://doi.org/ 10.4103/0972-6748.62274

Baron, A. (1999). Statistical inference in behavior analysis: Friend or foe? *The Behavior Analyst, 22*(2), 83–85.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. arXiv *[stat.CO]*. Retrieved from http://arxiv.org/abs/1406.5823

Boisgontier, M. P., & Cheval, B. (2016). The ANOVA to mixed model transition. *Neuroscience and Biobehavioral Reviews, 68*, 1004–1005. https://doi.org/ 10.1016/j.neubiorev.2016.05.034

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology &*

Evolution, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst*, *22*(2), 87–92.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378-400.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. https://doi.org/10.3389/fpsyg.2014.00823

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification*, *32*(6), 828–839. https://doi.org/10.1177/0145445508318606

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification*, *41*(2), 323–339. https://doi.org/10.1177/0145445516673998

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. https://doi.org/

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes*, *54*(1-3), 137–154.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*(3), 387–406.

Fisher, W. W., & Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: lies, damn lies, and statistics." *Journal of School Psychology*, *52*(2), 243–248. https://doi.org/10.1016/j.jsp.2014.01.001

Friedel, J. E., DeHart, W. B., Frye, C. C. J., Rung, J. M., & Odum, A. L. (2016). Discounting of qualitatively different delayed health outcomes in current and never smokers. *Experimental and Clinical Psychopharmacology*, *24*(1), 18–29. https://doi.org/10.1037/pha0000062

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

IBM SPSS Statistics Overview. (n.d.). Retrieved August 24, 2018, from https://www.ibm.com/products/spss-statistics

Killeen, P. R. (2018). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-018-0171-8

Kirkpatrick, K., Marshall, A. T., Steele, C. C., & Peterson, J. R. (2018). Resurrecting the individual in behavioral analysis: Using mixed effects models to address nonsystematic discounting data. *Behavior Analysis: Research and Practice*, *18*(3), 219–238. https://doi.org/10.1037/bar0000103

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, *6*(2), 151–157.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1071/AH17219

Lenth, R. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(2).

Mani, S., Sharma, S., & Singh, D. K. A. (2018). Web plot digitizer software: Can it be used to measure neck posture in clinical practice? *Asian Journal of Pharmaceutical and Clinical Research*, *11*(Special2), 86–87.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314.

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. https://doi.org/10.1016/j.jsp.2013.11.003

NIH (n.d.). Retrieved December 23, 2018, from https://grants.nih.gov/grants/how-to-apply-application-guide/forms-e/general/g.500-phs-human-subjects-and-clinical-trials-information.htm

Nugent, W. R. (1996). Integrating single-case and group-comparison designs for evaluation research. *The Journal of Applied Behavioral Science*, *32*(2), 209–226.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Parsonson, B. S. (1999). Fine grained analysis of visual data. *Journal of Organizational Behavior Management*, *18* (4), 47–51.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*(2), 109–116.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reed, D. D., & Azulay, R. L. (2011). A Microsoft Excel(®) 2010 based tool for calculating interobserver agreement. *Behavior Analysis in Practice*, *4*(2), 45–52.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*(2), 149–178. https://doi.org/10.1016/j.jsp.2013.11.004

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston: Authors Cooperative.

StataCorp, L. P., & Others. (2007). Stata data analysis and statistical software. https://www.stata.com/

Young, M. E. (2017). Discounting: A practical guide to multilevel analysis of indifference data. *Journal of the Experimental Analysis of Behavior, 108*(1), 97–112. https://doi.org/10.1002/jeab.265

Young, M. E. (2018a). A place for statistics in behavior analysis. *Behavior Analysis: Research and Practice, 18*(2), 193–202. https://doi.org/10.1037/bar0000099

Young, M. E. (2018b). Discounting: A practical guide to multilevel analysis of choice data. *Journal of the Experimental Analysis of Behavior, 109*(2), 293–312. https://doi.org/10.1002/jeab.316

Received: August 28, 2018
Final Acceptance: January 23, 2019
Editor in Chief: Michael Young
Associate Editor: Sarah Cowie