Behavior Analysis: Research and Practice
2018, Vol. 18, No. 2, 193–202

© 2018 American Psychological Association
2372-9414/18/$12.00 http://dx.doi.org/10.1037/bar0000099

A Place for Statistics in Behavior Analysis

Michael E. Young
Kansas State University

Behavior analysis has had an uneasy relationship with statistics. A historical reliance on visual inspection has slowly given way to null hypothesis significance testing and quantitative modeling, but this path has been fraught with missteps. There are challenges with reducing variability while also faithfully representing the variability that remains, and aggregation of outcome variables can undermine these goals. This manuscript highlights the shortcomings of human perception and judgment and the need to use modern statistical analysis to describe behavior and the uncertainty in these descriptions; statistical hypothesis testing based on a threshold should not be used as a substitute for human inference. A case is made for the increased use of generalized linear modeling, multilevel modeling, and model comparison as well as a need for stronger statistical training in behavior analysis programs.

Keywords: statistics, research methods, behavior analysis

Statistics and behavior analysis have had an uneasy relationship. The originator of the field of behavior analysis, B. F. Skinner, eschewed statistical analysis for a more experimental approach to controlling extraneous sources of variability and used the eye to discern the effects of interventions especially in cumulative records (Skinner, 1938). This tradition has influenced behavior analysis throughout much of its history. More recently, however, the emergence of the quantitative analysis of behavior introduced mathematical modeling to the field, and inferential statistical tests in the form of null-hypothesis statistical tests have slowly crept into behavior analysis for a variety of reasons (Shull, 1999). The historical skepticism about statistics has led to its de-emphasis in many behavior analytic training programs, and, when applied, statistical analysis in behavior analysis has tended to lag emerging trends in other fields of experimental psychology.

My goal in this article is to consider limitations in how data have been treated in behavior analysis to motivate a proper use of statistics that elucidates the issues of interest to the field. I will then progress to some recommendations on how statistics can be applied to enrich an understanding of behavior when they are used appropriately. Finally, I will close with a call for improved statistical training in behavior analysis programs.

Skepticism Toward Statistics in Behavior Analysis

The concerns about statistics in behavior analysis go back to its foundation. B. F. Skinner (1956) wanted to use experimental control to "reduce the troublesome variability" (p. 229) in behavior and thus "eliminate in advance of measurement the individual differences which obscure the difference under analysis" (p. 229). He sought to reduce scientific practice to the application of looking, just like a biologist can simply look at a specimen by using a microscope. It is important to recognize that his statements and perspectives developed during the early years of statistical analysis in psychology. Indeed, one of his dislikes for statistics was the tendency for statisticians of his day to push for designs that yield data conducive to analysis using the limited techniques that were then available. I am in complete agreement with Skinner in his opposition to those statisticians—the nature of the problem should dictate the tools that are used rather than vice versa. However, I believe that a

Correspondence concerning this article should be addressed to Michael E. Young, Department of Psychological

Sciences, 492 Bluemont Hall, Kansas State University,
Manhattan, KS 66506. E-mail: michaelyoung@ksu.edu

strict adherence to Skinner´s experimental practices can restrict our understanding of behavior in the same way that limitations in statistical analysis in the early 20th century restricted experimental design.

There are problems that emerge when analysis is done by simply looking at the results—the approach assumes that the perceiver can accurately discern the nature of what is being observed. Any visual stimulus comprises a wide range of features that must be properly integrated to produce an accurate judgment of the stimulus. When observing data, the presence of variability due to a variety of sources complicates judgment. Methodologists like Skinner (1938) and Murray Sidman (1960) fully appreciated this issue and thus strongly advocated an experimental approach to reduce this troublesome variability. Throughout the history of behavior analysis, however, the problem of perceiving differences in the face of variability has meant that statistical aggregation has subtly crept in. Any form of aggregation helps to smooth the data thus making patterns easier to see, but it also can hide trial-by-trial variability and sample size information. Variability, sample size, and other data characteristics should be informing a scientist's inferences but cannot do so when they are not faithfully represented in a figure. All aspects of data are critical to proper inference, and thus the loss of information before analysis, whether that analysis is visual or statistical, has the potential to lead to misguided conclusions (DeProspero & Cohen, 1979; Jones, Weinrott, & Vaught, 1978).

People do not have a history of being able to correctly integrate the many data features necessary to proper inference. In addition to intrasubject variability of the sort common in behavior analysis, proper inference must integrate intersubject variability, sample size, distributional properties of the response variable (e.g., is it normal, multimodal, skewed, or binomial), the magnitude and nature of relationships (e.g., linear vs. logarithmic), moderating variables, and data dependencies. These data characteristics are critical to proper inference.

Unfortunately, the field of judgment and decision making has documented that people are prone to a wide range of biases when simply looking at data. The ability of statistical and mathematical models to outperform experts in a variety of fields is testimony to limitations in people's ability to accurately discern relationships (Grove, Zald, Lebow, Snitz, & Nelson, 2000). Sedlmeier (1999) provided a nice treatise of the many biases that emerge when people are making statistical judgments even when provided with visual summaries of relationships rather then merely experiencing them as an expert often must do. In sum, people are simply not good intuitive statisticians. This observation leads to an appropriate skepticism when a scientist is faced with a presentation or manuscript with conclusions that rest heavily on visual inspection.

That said, too often statistics have been used improperly, which can lead to obfuscation, incorrect conclusions, useless inferences, and scientific malpractice (e.g., Bakker & Wicherts, 2011; Cohen, 1994; Simmons, Nelson, & Simonsohn, 2011). The *Behavior Analyst* published a special section on statistical inference in 1999, and the contributors to this section presented a wide range of well-documented problems with the misapplication of statistics (e.g., Baron, 1999; Branch, 1999; Perone, 1999). These include problems with null hypothesis significance testing, improper inferences based on the *p* value, the focus on whether there is a difference rather than the magnitude of that difference, and the loss of information about individual performance when averaging across members of a group. Many of these contributors lamented the large increase in statistical inference in publications involving behavior analysis. However, some of the contributors sounded a more cautious note (e.g., Ator, 1999; Crosbie, 1999; Shull, 1999). The one conclusion that all of these authors agreed on is that the misapplication of statistical analysis is not justifiable. Much of the criticism is appropriately directed at statistical inference leading to a decision (was there an effect or not) that was driven by a poorly understood *p* value. Furthermore, many of the comments were specific to methods that aggregate performance across individuals, which creates problems that should be (but alas, are often not) well understood (Anderson & Tweney, 1997; Ashby, Maddox, & Lee, 1994; Guthrie & Horton, 1946; Restle, 1965).

Issues in Behavior Analysis

Next I will consider a number of practices common to the behavior analytic approach and some possible consequences of their use. These people's ability to accurately discern relationships as a result of extended observation

some possible consequences of their use. These practices include training to asymptote in order to reduce variability, control of extraneous sources of variability, aggregating to create stable baselines or to ease visualization, and fitting highly parameterized models to individuals.

Training to Asymptote

One method of creating a stable baseline to make it easier to visualize the effect of an intervention is to train an organism to asymptote (i.e., to "steady-state"). Having a stable baseline is central to the philosophical foundations of the experimental analysis of behavior because it allows the experimenter to discern the immediate consequences of an intervention, thus supporting causal inference. Although replication strengthens these conclusions by either returning behavior to baseline and reintroducing the intervention (e.g., an ABAB design) or intervening at different times for different individuals (a multiple baseline across subjects design), very little such replication is necessary given the small degree of variation purportedly present in the baseline. There are two issues that arise with this approach: achieving stability and consideration of behavior in transition.

Achieving stability requires establishing a criterion against which stability is assessed. In published articles, it is not immediately apparent how any particular criterion was chosen. But, as Sidman (1960) observed, "there must be a considerable amount of experience and intuition involved in the selection of an appropriate criterion" (p. 259). A researcher can use published studies, and prior and current observation of a subject, to decide when stability has occurred, and this criterion may actually vary from subject to subject. These researcher degrees of freedom, however, can produce increases in false alarm rates in the data collection process just as they can during statistical analysis (Simmons et al., 2011).

By definition, behavior in transition is encountered when a subject's behavior is not stable. Transitional behavior has been central to a wide range of questions of interest involving learning and the latency to full effect of an intervention. Sidman (1960) dedicated a chapter in *Tactics of Scientific Research* to the consideration of transitions after noting that "transition states are of potential interest as important behavioral phenomena in their own right" (p. serious consideration of transitional behavior, and scientists examining animal behavior have often found that mathematical modeling of transitional behavior has been especially helpful (e.g., Banna & Newland, 2009; Gallistel, Mark, King, & Latham, 2001; Killeen & Fetterman, 1993; Myerson & Hale, 1988).

Control of Extraneous Sources of Variability

When extraneous variables are held constant, it raises the question of whether the results of a study extend beyond the particular conditions under which a phenomenon was observed. In an era when replication has been shown to be a serious problem (Open Science Collaboration, 2015), controlling variability can undermine replication across contexts and laboratories. In a response to the problem of replicating animal experiments, Richter, Garner, Auer, Kunert, and Würbel (2010) considered whether the highly standardized designs common to animal studies (holding constant variables like age, sex, body weight, bedding, and lighting conditions) might be producing results that are specific to the set of conditions under which they were produced. The researchers found that standardization increased the sensitivity of a manipulation but at the cost of reproducibility across what appear to be inconsequential changes to the conditions (e.g., 8- or 14-week-old mice, morning or afternoon testing, normal vs. dim lighting, and the identity of the experimenter). While it is critical to the internal validity of an experiment that extraneous factors do not systematically vary across experimental conditions, thus introducing confounds, tight control of all sources of variation in order to ease visualization of causal effects might have the unfortunate side effect of limiting the generalizability of the results.

Of course, part of the issue may be with the scientists who infer or claim that they are establishing general phenomena rather than with the desire for experimental control. For example, it is less impactful to conclude that 16-week-old rats are subject to hyperbolic discounting of food pellets in a particular bar press task than to make general claims that this principle holds across species, reward types, delay durations, and response operanda. It is only through the systematic study of variations that a behavioral phenomena in their own right" (p. 263). Sidman noted the many challenges to the

through the systematic study of variations that a generalizable principle can be established. Naturally, this type of study requires considerable resources, and most laboratories only have a finite number of identical Skinner boxes in which to run their experiments. Furthermore, variation may reveal that a phenomenon is not very robust, thus undermining its impact; when successful publication is necessary for career advancement, there exist disincentives to systematic variation (Nuzzo, 2015).

Aggregating to Create Stable Baselines or to Ease Visualization

Researchers are often using statistical aggregation when they are plotting their data for visual inspection. For example, any plot showing response rate as the dependent variable has used aggregation because the experimenter needed to choose a window of time to derive that rate (e.g., responses per minute or hour). Also, any plot showing a percentage as the dependent variable used aggregation by choosing a block of behavior on which to compute that percentage (e.g., percentage of correct responses in each block of 10 or 100). Aggregating can hide two sources of information, sample size and variability, both of which are critical for proper inference. To avoid aggregation, I recommend the statistical analysis of interresponse times (IRTs) rather than response rate (cf. D. S. Blough, 1963; P. M. Blough & D. S. Blough, 1968) and the use of repeated measures logistic regression for choice data (cf. Dixon, 2008).

To illustrate the limitations of visual analysis, I simulated data from a simple design involving two conditions for a single subject. I created IRTs randomly drawn from a Gamma distribution because this procedure produces the type of non-negative skewed IRTs common to studies of schedules of reinforcement. In the first task simulation, IRTs were assumed to be constant across 500 responses, to mildly vary across responses, and to be higher for one condition than for the other (population mean of 3.0 vs. 2.0). In the second simulation, the same situation held except that the IRTs were more variable. In the third simulation, IRTs decreased linearly (in log-transformed space) as a function of response number with the two groups differing in the initial IRT and the rate of IRT change. In the fourth simulation, IRTs decreased in a curvilin-

early in training than they did later (in log-transformed space), again with a group difference in the initial IRT and rate of IRT change. The need to qualify statements regarding behavioral change to log-transformed space illustrates one of the complexities when dealing with rate-related outcomes (and, incidentally, in proportion or percentage data): The presence of a data floor or ceiling (e.g., a minimum IRT of zero) will distort the assessment of a variable's effect if not taken into consideration (Bartlett, 1947; Dixon, 2008; Lo & Andrews, 2015).

The top row of Figure 1 plots the data for these four task simulations as cumulative response records. Cumulative response records are especially useful for depicting response rate differences despite noise in the data, as illustrated by the first and second plots. The third and fourth plots are familiar to behavior analysts in that they depict an increase in response rate across time as is common in fixed interval schedules when the scheduled reinforcement time is approaching. Cumulative records have a few shortcomings that should be noted. First, they require that the reader be able to attend to slope differences rather than level differences. This feature makes it difficult to numerically compare two conditions. For example, it would require extra effort to derive rate estimates for the two lines in the first two plots so that they can be compared. Second, cumulative records make it difficult to evaluate nonlinear changes in the response rate. For example, although an experienced analyst knows that the rate of change is increasing in the third and fourth plots, the rate of that increase and whether it is linear or curvilinear is not evident.

The second row of Figure 1 depicts the IRTs for the 500 responses that produced the cumulative records in the first row. A few observations are immediately apparent. First, the greater IRT variability in the second task is much more obvious. Second, the changes in IRTs across time are easier to discern—a linear change in the third task versus a curvilinear change in the fourth task. Third, it is much more difficult to observe the A versus B difference in the second task with high IRT variability; this difference was much easier to recognize in the cumulative record. Finally, the variability in IRTs is decreasing across time in the third and fourth simulation, IRTs decreased in a curvilinear fashion such that they changed more rapidly

IRTs is decreasing across time in the third and fourth simulated tasks.

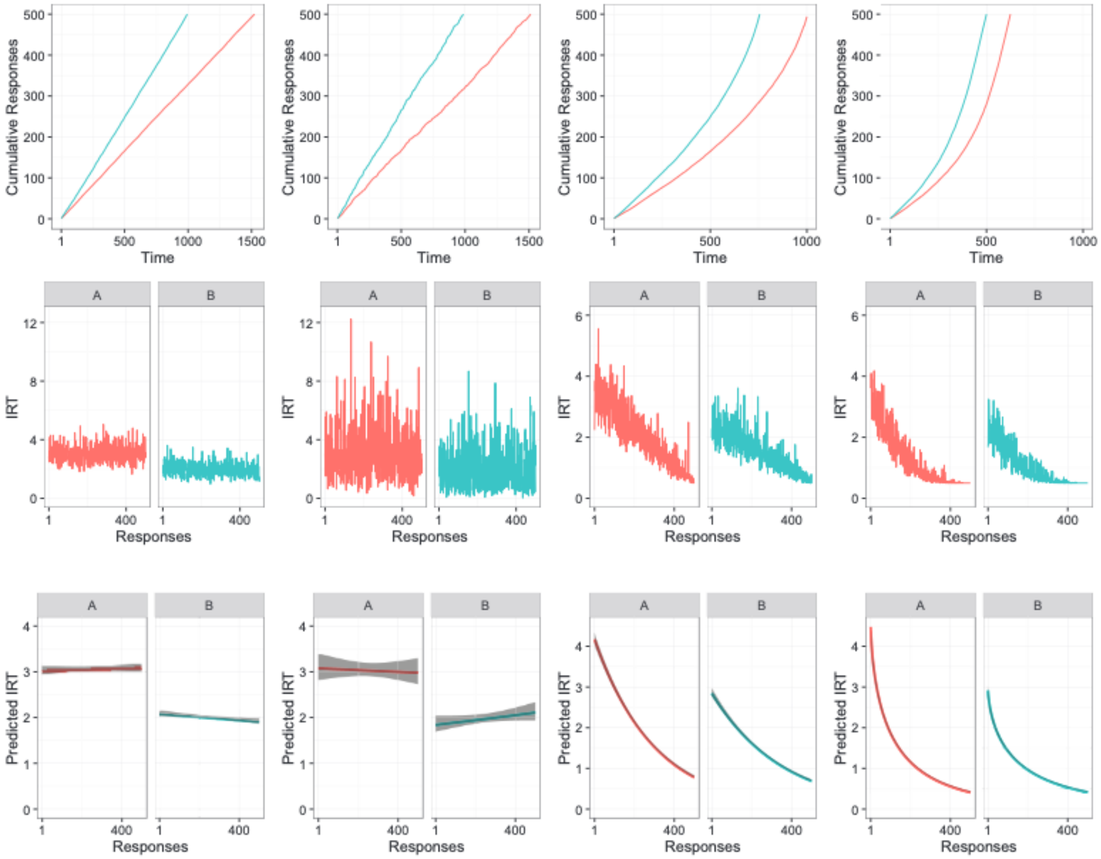

Figure 1. Three representations (top to bottom: cumulative record, IRT, and statistically-estimated IRT) of four behavioral patterns (left to right: constant clean, constant noisy, linearly decreasing, and curvilinear decreasing). See text for more explanation.

By plotting individual IRTs rather than aggregated response rates, the full range of variability can be considered not only visually but also by a statistical analysis of these data. An alternative is to create rates by plotting and analyzing 1/IRT, but my experience has revealed that inverted rates are much more skewed than IRTs and inversion is more likely to produce outliers (when the IRT is very small). Regardless, an appropriate statistical analysis of individual IRTs will be based on every response, thus increasing sample size, incorporating variability in the IRTs, and modeling the IRT distribution (often well-approximated by Gamma distributions). A statistical analysis will also generate data descriptions that will prove useful to the behavior analyst by focusing attention on the precise estimation of

The third row of Figure 1 plots statistical fits to these data (note that the *y*-axis range has been changed to make it easier to compare the fits). Each fit includes a 95% confidence error ribbon that depicts the uncertainty in the model fit—very small for the low variability IRTs in the first, third, and fourth tasks, and moderate for the high variability IRTs in the second task. Often when a researcher plots aggregates like response rates and percentages, measures of uncertainty like these error ribbons are missing.

As in any statistical analysis, the accuracy of a fit is heavily dependent on the use of an appropriate statistical model. In this case, I used a generalized linear model to perform Gamma regression (i.e., I assumed that the IRTs followed a Gamma distribution). For the third and fourth tasks, I focusing attention on the precise estimation of magnitudes and their uncertainties.

distribution). For the third and fourth tasks, I tested multiple models in which the relationship between response number and IRT was either assumed to be linear (in log-transformed space) or to follow a power function.

An appropriate statistical analysis readily revealed the condition differences for the first two tasks. The 95% confidence intervals of the predicted IRT for the two conditions in the first task were [2.99, 3.10] and [1.95, 2.02] and in the second task were [2.86, 3.20] and [1.86, 2.08]; both intervals encompassed the population values of 3.0 and 2.0 but with more uncertainty for the estimate based on the noisier data. Although the condition difference was statistically significant for both conditions, the key advantage of the statistical analysis was the ability to precisely estimate each condition's average IRT and its uncertainty, not merely to determine whether they were different. Furthermore, neither analysis indicated that the IRTs were systematically changing across time (i.e., the response slope was about zero). The analyses of tasks three and four revealed that the A and B conditions generated different IRT curves, but more critically the analyses were also able to estimate intercepts and slopes, and to evaluate the possible presence of nonlinear change. Analysis of the third task did not reveal a nonlinearity (the appearance of curvature in Figure 1 is due to the Gamma distribution of IRTs), whereas the analysis of the fourth task revealed curvature well-approximated by a power function.

This example focused on a single subject and incorporated intrasubject variability, sample size, and a non-normal IRT distribution. But, proper inference must integrate intersubject variability, the number of subjects, and differential data dependencies (e.g., values collected from the same subject in the same condition are expected to be more highly correlated with one another than values collected from different subjects or in a different condition). Visual inspection may be fine for identifying simple relationships, but when aggregation occurs to ease that inspection, necessary information has been lost to both the eye and to any analysis of the aggregated data.

Fitting Highly Parameterized Models to Individuals

The quantitative analysis of behavior has had a long history in behavior analysis. The generalized matching law (Herrnstein, 1961), hyperbolic discounting curve (Rachlin, Raineri, &

(White & Wixted, 2010) have all had considerable success at modeling behavior. Unfortunately, many of these and other modeling approaches have been applied to aggregated data like response rates, the proportion of larger later responses, or the probability of a false alarm. Statistical analysis of individual responses in the behavioral sciences rarely generates R^2 values over 0.9 due to the considerable variation in the behavior of complex biological organisms. Thus, when I observe a remarkably good fit for these types of quantitative models even at the individual level, then it is highly likely that some type of variability was not modeled.

Because aggregation significantly reduces the number of observations being used in a model fit, only simple models can be fit at the individual level. Frequently there is not enough data to reliably estimate complex models when the only data being used are those generated from a single individual, and this problem is exacerbated when the data are aggregated within-subject. For example, two-parameter hyperbolic discounting functions have proven to be much more difficult to fit when applied to individual-subject aggregated indifference point data (Young, 2017). There are statistical tools that can stabilize the estimation of model parameters for more complex models, but these approaches rely on some combination of (a) using disaggregated data to increase sample size and (b) making more educated estimates by considering the behavior of other subjects in the experiment.

A Proper Place for Statistics in Behavior Analysis

"Statistical theory has provided us with a toolbox with effective instruments, which requires judgment about when it is right to use them" (Gigerenzer, 2004). Statistics is a tool—it is the application of the tool that matters. When the tool is severely limited as it was during Skinner's era, its proper use will likewise be limited. When a tool is inappropriately used only because of its ease-of-use or familiarity, the outcome has the potential to produce an inappropriate description of the data.

Null hypothesis significance testing came to dominate experimental psychology, and it has been properly criticized by many (Cohen, 1994; Fidler, Geoff, Burgman, & Thomason, 2004;

bolic discounting curve (Rachlin, Raineri, & Cross, 1991), and signal detection theory

Fidler, Geoff, Burgman, & Thomason, 2004, Loftus, 1996; Meehl, 1978). Once it established a foothold in statistical software and graduate training programs, it was both easy to use and familiar. Instead of focusing on testing of the null hypothesis, statistics can provide a foundation for the accurate assessment of quantities of interest and an assessment of the uncertainty in these estimates. Fields like physics, chemistry, and biology use statistics for precisely this purpose. Pick up any issue of *Nature* or *Science* to see numerous examples of the use of sophisticated statistical techniques harnessed to illuminate relationships in physics, biology, chemistry, and neuroscience *inter alia*. Although an adoption of statistics in behavior analysis is problematic if it resorts to an overreliance on null hypothesis significance testing and group analysis, statistical analysis is very helpful when used to properly estimate the rate of behavior, the variability in behavior, the rate of behavioral change, the size of the differences in these quantities across conditions or groups, or the degree of impact of another variable on these quantities, at both the individual and group levels. The challenge is analyzing data properly so as to be confident in these estimates.

Like any summary, statistical descriptions can be misleading if an analysis is misapplied. Obvious examples include the need to recognize the impact of an outlier, how correlations assume that a relationship between two variables is linear, and the unfortunate effects of averaging multimodal data. But inexperienced data analysts may not recognize the consequences of basing an analysis on statistical aggregates, the presence of heterogeneity of variance, overfitting when using a complex model, and differential dependencies among data values (i.e., the nature of the variance/covariance structure).

I believe that there are three techniques with which behavior analysts should gain greater familiarity: generalized linear modeling, multilevel modeling, and Bayesian analysis. Although there are a number of other tools that can help enlighten data of interest to a behavior analyst, these three techniques can address the most egregious issues plaguing contemporary use of statistics in all fields of psychology that rely on repeated measures designs.

Generalized linear modeling offers two principal benefits: the ability to seamlessly integrate categorical and continuous predictors of behavior and the proper analysis of outcome variables that are not well-described by a normal curve. All modern statistical software includes generalized linear modeling. By treating variables like trial and session as continuous predictors of behavior, the analyst shifts from a focus on identifying whether there are differences between the levels of these variables to a focus on the slope of the relationship. Being able to estimate the slope of a function, how other variables alter this slope, and the type of relationship present (linear, logarithmic, power, etc.) offers much greater utility that merely cataloging categorical differences. Furthermore, generalized linear modeling is an extension of GLM by adding the capability to correctly analyze data that violate GLM's distributional assumptions. Specifically, generalized linear models allow the analyst to specify a binomial distribution (for binomial choice data), a Poisson distribution (for count data), and Gamma distributions (for latency data). The distributions of each of these three types of data are problematic for GLM and can lead to substantial model mis-specification.

Multilevel modeling is now available in all major statistical programs, and generalized multilevel modeling is available in many of them. Multilevel modeling (also known as mixed effects modeling) is a great tool for behavior analysts because it estimates model parameters at the individual and group levels simultaneously (Boisgontier & Cheval, 2016; Bolker et al., 2009; Gelman & Hill, 2006; Gueorguieva & Krystal, 2004; Young, 2017; Young, Clark, Goffus, & Hoane, 2009). Because multilevel modeling can be used with both categorical and continuous predictors, it empowers the researcher to break from the traditional restriction of using categorical predictors as required by repeated measures ANOVA. Furthermore, the approach estimates parameter values (e.g., the rate of change across trials, days, sessions, or the magnitude of the difference between conditions) both at the group level (called the fixed effect of the variable) as well as at the individual level (by incorporating the random effect of the variable). When the outcome variable is binomial, generalized multilevel modeling can be used, thus making repeated measures logistic regression possible. Indeed, the scientist is no longer bound to the normality assumption in repeated measures analysis, because multiple categorical and continuous predictors of behavior (like the general linear model, GLM), and

repeated measures analysis, because multiple types of distributions can be specified.

Finally, Bayesian statistics encourages the data analyst to move beyond the traditional *p* value that focuses attention on the probability of the data given the null hypothesis, P(D | H_0), and toward thinking about the relative probability of multiple hypotheses, like P(H_1 | D) versus P(H_2 | D). By focusing on model comparison, Platt's (1964) method of strong inference is encouraged, and the behavior analyst can quantify the strength of the evidence in favor of one hypothesis versus another hypothesis where neither one has to be the null. For example, in my analyses of the third and fourth tasks of Figure 1, I was able to compute likelihood ratios and Bayes factors to evaluate the evidence in favor of a linear versus a power relationship between response number and IRT—the uninformative null hypothesis was not seriously considered. Although Bayesian statistical analysis is not easily performed in many statistical packages, familiarity with the technique and its logic will hone the thinking of data analysts.

Thus, the modern behavior analyst is no longer faced with the choice between the limitations of visual inspection and the confines of null hypothesis significance testing. The key challenge is equipping scientists to ensure that these new statistical tools are used properly rather than merely representing a new way to make the same mistakes.

A Call for Stronger Statistical Training for Behavior Analysts

As the sophistication of statistical tools has increased, the knowledge of their availability and proper use has lagged. Indeed, contemporary approaches to data analysis are heavily influenced by historical approaches. For example, the early adoption of the analysis of variance (ANOVA) for categorical predictors has dominated experimental psychology for so long that researchers faced with continuous predictors will make them into categorical ones to support their use of ANOVA (Young, 2016). For the reasons cited above, I encourage programs in behavior analysis to implore their students to obtain stronger statistical expertise. I certainly am not the first to make this plea. In 1999, Nancy Ator was one of a few contributors to a special issue of *The Behavior Analyst* mak-

> When behavior analysts do not understand design and statistical methods well enough, we become unduly subject to the preconceived notions of reviewers, editors, colleagues, and department chairs . . . we should be able to review manuscripts well enough to understand whether the statistics included are appropriate and appropriately described. (Ator, 1999, pp. 95–96)

In many psychology graduate programs, there is little or no expertise available among the faculty to train students in modern statistical techniques. During my years in graduate school in the early 1990s, I received no training in generalized linear modeling, multilevel modeling, Bayesian approaches, or a number of other techniques; in each case my knowledge was obtained by reading, hands-on experience, and the kind responses of people who responded to my email queries. As I gained greater confidence in my own expertise, these topics were gradually integrated into my graduate statistics courses. Thus, graduate programs need to determine whether training exists elsewhere on campus, whether they can hire the expertise, or if one or more current faculty will need to obtain expertise in order to pass it along to students and colleagues. Although a text like Huitema's (2011) is a solid introduction to statistics for behavior analysts that covers the necessary precursors, students and instructors will need to work harder to generalize existing texts that cover multilevel modeling (e.g., Gelman & Hill, 2006) and Bayesian approaches (e.g., Kruschke, 2014) to behavior analysis.

Until such expertise is more widespread, behavioral scientists will encounter barriers in the proper analysis of their data, in the review process when new techniques are treated with skepticism, and in impact when readers fail to understand the results produced by these techniques. Regardless, the ability to describe data more accurately and with a valid measure of uncertainty is an ability worth pursuing. Sidman (1960) recognized the difference between statistical evaluation and statistical description and had no trouble with the latter. The professed advantage of using statistical evaluation (via the *p* value) as a substitute for human evaluation of an observed quantity is that it removes human judgment from the decision-making process. Nevertheless, this removal of human judgment is an illusion because all research relies on the ethics, judgment, and intelligence of the experto a special issue of *The Behavior Analyst* making the same appeal:

culties, judgment, and intelligence of the experimenter. A scientist can always engage in questionable research practices regardless of the use of visual inspection, null hypothesis significance testing, Bayesian approaches, or statistical summary. An honest description of results is critical to good science, whether this description entails intercepts, slopes, other parameter estimates, R^2, likelihood ratios, or Bayes factors.

A fine-grained analysis of behavior is necessary to identify cause-effect relations. Behavior analysts should be at the forefront of the proper analysis of repeated measures data given their commitment to within-subject designs and a focus on replication. Rather than adopting tired statistical approaches or retreating to visual inspection that is prone to cognitive and visual illusions, the field of the experimental analysis of behavior can bring rigor in research design *and* data analysis to the scientific study of behavior.

References

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition, 25,* 724–730. http://dx.doi.org/10.3758/BF03211315

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5,* 144–151. http://dx.doi.org/10.1111/j.1467-9280.1994.tb00651.x

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *The Behavior Analyst, 22,* 93–97. http://dx.doi.org/10.1007/BF03391985

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. http://dx.doi.org/10.3758/s13428-011-0089-5

Banna, K. M., & Newland, M. C. (2009). Within-session transitions in choice: A structural and quantitative analysis. *Journal of the Experimental Analysis of Behavior, 91,* 319–335. http://dx.doi.org/10.1901/jeab.2009.91-319

Baron, A. (1999). Statistical inference in behavior analysis: Friend or foe? *The Behavior Analyst, 22,* 83–85. http://dx.doi.org/10.1007/BF03391983

Bartlett, M. S. (1947). The use of transformations. *Biometrics, 3,* 39–52. http://dx.doi.org/10.2307/3001536

Blough, D. S. (1963). Interresponse time as a function of continuous variables: A new method and some data. *Journal of the Experimental Analysis of Behavior, 6,* 237–246. http://dx.doi.org/10.1901/

Blough, P. M., & Blough, D. S. (1968). The distribution of interresponse times in the pigeon during variable-interval reinforcement. *Journal of the Experimental Analysis of Behavior, 11,* 23–27. http://dx.doi.org/10.1901/jeab.1968.11-23

Boisgontier, M. P., & Cheval, B. (2016). The anova to mixed model transition. *Neuroscience and Biobehavioral Reviews, 68,* 1004–1005. http://dx.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. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24,* 127–135. http://dx.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,* 87–92. http://dx.doi.org/10.1007/BF03391984

Cohen, J. (1994). The earth is round (p <. 05). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Crosbie, J. (1999). Statistical inference in behavior analysis: Useful friend. *The Behavior Analyst, 22,* 105–108. http://dx.doi.org/10.1007/BF03391987

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579. http://dx.doi.org/10.1901/jaba.1979.12-573

Dixon, P. (2008). Models of accuracy in repeated-measures design. *Journal of Memory and Language, 59,* 447–456. http://dx.doi.org/10.1016/j.jml.2007.11.004

Fidler, F., Geoff, C., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics, 33,* 615–630. http://dx.doi.org/10.1016/j.socec.2004.09.035

Gallistel, C. R., Mark, T. A., King, A. P., & Latham, P. E. (2001). The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes, 27,* 354–372. http://dx.doi.org/10.1037/0097-7403.27.4.354

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790942

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33,* 587–606. http://dx.doi.org/10.1016/j.socec.2004.09.033

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Gueorguieva, R., & Krystal, J. H. (2004). Move over

Behavior, 6, 237–246. http://dx.doi.org/10.1901/jeab.1963.6-237

Gueorguieva, R., & Krystal, J. H. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the *Archives of General Psychiatry*. *Archives of General Psychiatry, 61,* 310–317. http://dx.doi.org/10.1001/archpsyc.61.3.310

Guthrie, E. R., & Horton, G. P. (1946). *Cats in a puzzle box*. New York, NY: Rinehart.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4,* 267–272. http://dx.doi.org/10.1901/jeab.1961.4-267

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781118067475

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277–283. http://dx.doi.org/10.1901/jaba.1978.11-277

Killeen, P. R., & Fetterman, J. G. (1993). The behavioral theory of timing: Transition analyses. *Journal of the Experimental Analysis of Behavior, 59,* 411–422. http://dx.doi.org/10.1901/jeab.1993.59-411

Kruschke, J. K. (2014). *Doing Bayesian data analysis*. Waltham, MA: Elsevier.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6,* 1171. http://dx.doi.org/10.3389/fpsyg.2015.01171

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171. http://dx.doi.org/10.1111/1467-8721.ep11512376

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

Myerson, J., & Hale, S. (1988). Choice in transition: A comparison of melioration and the kinetic model. *Journal of the Experimental Analysis of Behavior, 49,* 291–302. http://dx.doi.org/10.1901/jeab.1988.49-291

Nuzzo, R. (2015). How scientists fool themselves—and how they can stop. *Nature, 526,* 182–185. http://dx.doi.org/10.1038/526182a

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22,* 109–116. http://dx.doi.org/10

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146,* 347–353. http://dx.doi.org/10.1126/science.146.3642.347

Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the Experimental Analysis of Behavior, 55,* 233–244. http://dx.doi.org/10.1901/jeab.1991.55-233

Restle, F. (1965). Significance of all-or-none learning. *Psychological Bulletin, 64,* 313–325. http://dx.doi.org/10.1037/h0022536

Richter, S. H., Garner, J. P., Auer, C., Kunert, J., & Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. *Nature Methods, 7,* 167–168. http://dx.doi.org/10.1038/nmeth0310-167

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical applications*. Mahwah, NJ: Erlbaum.

Shull, R. L. (1999). Statistical inference in behavior analysis: Discussant's remarks. *The Behavior Analyst, 22,* 117–121. http://dx.doi.org/10.1007/BF03391989

Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Skinner, B. F. (1938). *The behavior of organisms*. New York, NY: Appleton-Century-Crofts.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11,* 221–233. http://dx.doi.org/10.1037/h0047662

White, K. G., & Wixted, J. T. (2010). Psychophysics of remembering: To bias or not to bias. *Journal of the Experimental Analysis of Behavior, 94,* 83–94. http://dx.doi.org/10.1901/jeab.2010.94-83

Young, M. E. (2016). The problem with categorical thinking by psychologists. *Behavioural Processes, 123,* 43–53. http://dx.doi.org/10.1016/j.beproc.2015.09.009

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

Young, M. E., Clark, M. H., Goffus, A., & Hoane, M. R. (2009). Mixed effects modeling of Morris water maze data: Advantages and cautionary notes. *Learning and Motivation, 40,* 160–177. http://dx.doi.org/10.1016/j.lmot.2008.10.004

Received July 20, 2017
Revision received September 16, 2017

.1007/BF03391988

Accepted September 19, 2017 ■