2022

A Monte Carlo Simulation of Rat Choice Behavior with Interdependent Outcomes

Michelle A. Frankot
West Virginia University, mf0083@mix.wvu.edu

A Monte Carlo Simulation of Rat Choice Behavior with Interdependent Outcomes

Michelle A. Frankot

Dissertation submitted to Eberly College
at West Virginia University

In partial fulfillment of the requirements for the degree of

Ph.D. in
Psychology, concentration in Behavioral Neuroscience

Melissa Blank, Ph.D., Chair
Cole Vonder Haar, Ph.D.
Christa Lilly, Ph.D.
Michael Young, Ph.D.

Department of Psychology

Morgantown, West Virginia

2022

Keywords: animal behavior, choice, Monte Carlo, statistical models

ABSTRACT

A Monte Carlo Simulation of Rat Choice Behavior with Interdependent Outcomes

Michelle A. Frankot

Preclinical behavioral neuroscience often uses choice paradigms to capture psychiatric symptoms. In particular, the subfield of operant research produces nested datasets with many discrete choices in a session. The standard analytic practice is to aggregate choice into a continuous variable and analyze using ANOVA or linear regression. However, choice data often have multiple interdependent outcomes of interest, violating an assumption of general linear models. The aim of the current study was to quantify the accuracy of linear mixed-effects regression (LMER) for analyzing data from a 4-choice operant task called the Rodent Gambling Task (RGT), which measures decision-making in the context of various manipulations (e.g., brain injury). Prior analysis of RGT data from intact rats (Sham; n = 58) and brain-injured rats (TBI; n = 51) revealed five distinct decision-making phenotypes for this task. To generate datasets for parametric analysis, trial-level data was simulated using a Monte Carlo approach recapitulating those phenotypes. Population parameters were defined from existing data, and repeated sampling was conducted to generate 1000 datasets for four sample sizes (n = 6, 10, 14, 20) and four effect sizes (f = 0.0, 0.3, 0.4 and 0.5). Two LMER models were performed to compare TBI versus Sham across datasets: a full LMER where choice of all four outcomes was analyzed simultaneously, and a control LMER where choice of a single outcome was analyzed. The full LMER exceeded 75% false positives across all sample sizes, and the control LMER was underpowered to detect expected effects. These results suggest analyzing trial-level data in a mixed effects logistic regression will be necessary to accurately analyze RGT data. More broadly, these types of errors must be remedied to improve translation to clinical research.

Table of Contents

A Monte Carlo Simulation of Rat Choice Behavior with Interdependent Outcomes

Two major obstacles in basic scientific research are reproducibility and translation. Although the reproducibility crisis is a widespread phenomenon, the translation crisis is particularly pronounced in behavioral neuroscience. In fact, there are estimates that 90% of preclinical therapeutics in behavioral neuroscience fail to translate to humans (e.g., Garner, 2014). Translation has been exceptionally difficult in the subfield of traumatic brain injury (TBI) where successful treatments in rodents largely fail in clinical trials (Bragge et al., 2016). There are many complex factors that contribute to this disconnect, such as poor preclinical models, inherent species differences, miscommunication across disciplines, and lack of funding incentives.

Improper use of statistical tests may also contribute to the translation and reproducibility crises (Seyhan, 2019). In fact, a review of 125 peer-reviewed preclinical articles in the field of TBI and spinal cord injury found that 70% of papers contained an inappropriate statistical technique (e.g., incorrect post-hoc tests, incorrect use of parametric tests) (Burke, Whittemore, & Magnuson, 2013). Improper statistics at the preclinical level may produce findings that fail to translate to clinically-meaningful results. These types of data analytic issues are an excellent target for narrowing the translational gap because they do not require major scientific advancements to rectify. One particular analytic error that may create inaccuracies at the preclinical level is the use of parametric tests when core assumptions are violated. In the current study, we used a Monte Carlo approach to empirically identify any limitations in the traditional data analytic approach for a preclinical behavioral paradigm called the Rodent Gambling Task (RGT).

Violations of Parametric Statistics

The General Linear Model (GLM) is the foundation of many parametric statistical tests (e.g., *t*-test, ANOVA, linear regression) used to analyze preclinical data. The GLM models a linear relationship between variables in the form of [$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e$] where y is the dependent variable, x is the independent variable(s), the slope (β) is the strength of the relationship associated with each independent variable, β_0 is the intercept, and *e* represents the error term (Faraway, 2016). After fitting data to a linear model, the difference between the observed y-values and the predicted y-values are referred to as residuals. To analyze data using the GLM, the data must meet four explicit assumptions regarding the structure of the residuals: (1) the residuals must be independent of one another, (2) the residuals must be normally distributed (3), the residuals must have a mean of zero at all values of x, and (4) the residuals must have constant variance (i.e., homoscedasticity) (McCullagh, 1989).

However, real data tend to violate these underlying assumptions to varying degrees. Assumption violations can bias parameter estimates and the errors of those parameter estimates, which in turn can increase the risk of false positives (i.e., Type I error) and reduce the ability to detect true differences between groups (i.e., power, Type II error) (Erceg-Hurn & Mirosevich, 2008). Unfortunately, violations are often dismissed. When 30 PhD researchers in Psychology were given hypothetical datasets to analyze, fewer than 25% checked if assumptions were violated (Hoekstra, Kiers, & Johnson, 2012). Although the lack of awareness surrounding violations is concerning, assumption violations do not always have detrimental effects on data interpretation due to the robust nature of the GLM. Under certain conditions, data that violate assumptions (e.g., non-normal data void of substantial outliers) may still be analyzed appropriately using linear models (Knief & Forstmeier, 2020), although there is considerable

debate regarding the robustness of the GLM to violations (Bradley, 1978; Micceri, 1989). First, there is no operational definition of the "robustness" of the GLM. Second, there are no clear guidelines outlining what constitutes normal "enough" to analyze using the GLM. Thus, it is important to experimentally test whether violations in preclinical paradigms produce impactful errors. Specifically, we will consider violations of the assumption of independence of residuals, which occur frequently in preclinical choice paradigms.

Choice Paradigms in the TBI Field

Violations of the independence assumption often occur in behavioral neuroscience research due to the use of repeated measures. Fortunately, an easy remedy is to use an analysis that accounts for within-subject dependence, such as a repeated measures ANOVA or linear mixed model nested by subject. However, more problematic independence violations may be unavoidable in experiments that use choice paradigms, where outcomes are often interdependent by nature. When outcomes are interdependent, this increases the likelihood that residuals will be interdependent, thus violating GLM assumptions. In choice paradigms, interdependent outcomes may artificially inflate parameter estimates for the coefficients representing the effects of predictors in the model because when choice of one outcome shifts, it inherently changes choice of other outcomes. When the magnitude of difference across conditions is artificially inflated, false positives may be more likely to occur (i.e., researchers are more likely to find significant effects when a true population-level effect does not exist).

Choice paradigms are particularly relevant for modeling various psychiatric conditions (see Table 1 for a list of example paradigms). Psychiatric deficits, such as poor decision-making, are common after brain injury (Vaishnavi, Rao, & Fann, 2009), making preclinical TBI research susceptible to the independence violation because choice paradigms are often required. There are

two primary classifications of choice paradigms in preclinical TBI literature that violate the assumption of independence: (1) choice paradigms with two options and (2) choice paradigms with three or more options. It is important to distinguish between these because each requires a different solution. Specifically, violations in 2-choice paradigms can be easily remedied, while paradigms with additional choices require more consideration.

Table 1.

Common Preclinical Behavioral Paradigms with Interdependent Outcomes.

Task	**Choice Parameters**	**Pub-Med Hits**
Novel Object Test	2-Choice	4,638
Elevated Plus Maze	2-Choice	8,972
Conditioned Place Preference	2-Choice	5,192
Delay Discounting Task	2-Choice	1,115
Social Preference Test	2-Choice/3-Choice	8,114
Forced Swim	3-Choice	9,564
Rodent Gambling Task	4-Choice	159
Morris Water Maze (Quadrant Analysis)	4-Choice	11,957

Note: Search was conducted on October 12, 2021. Task names were used as search parameters.

Choice Paradigms with Two Options

In choice paradigms with two options, a preference score is typically used as a dependent variable. If preference is calculated as a ratio of one option divided by a total score, it does not violate the independence assumption. However, some researchers calculate preference as choice of one option compared against choice of another option (i.e., $\text{Preference} = \frac{\text{Option 1}}{\text{Option 2}}$). This approach is used in a 2-choice paradigm called the Novel Object Recognition (NOR) task

(Ennaceur & Delacour, 1988). During the task, rodents are presented with an object and given time to establish initial familiarity with it. Then, after some period of time (e.g., 1 hour), rodents are given time to choose to interact with either the familiar object or a novel object. The dependent variable is typically a preference index comparing time spent with each object, which is viewed as an indicator of the ability to discriminate between novel and familiar stimuli (Sivakumaran, Mackenzie, Callan, Ainge, & O'Connor, 2018). There are a variety of ways this index can be calculated (Antunes & Biala, 2012), some of which violate the independence assumption (but can be readily fixed). One common problematic calculation defines preference as $\frac{\%\text{ time with novel}}{\%\text{ time with familiar}}$ (Antunes & Biala, 2012; Broadbent, Gaskin, Squire, & Clark, 2009). If not transformed or accounted for, this approach artificially inflates the preference index; as time with the novel object increases, time with the familiar object inherently decreases, which amplifies the resulting calculation. Unfortunately, this is a common approach used by many current papers in the field (e.g., Bahceci, Anderson, Occelli Hanbury Brown, Zhou, & Arnold, 2020; Bruijnzeel et al., 2019; Hornoiu, Gigg, & Talmi, 2020; Munyon, Eakin, Sweet, & Miller, 2014).

However, this violation can be easily remedied with formulaic changes or minor transformations (e.g., percent or log transformation). For example, novel object preference could be calculated as $\frac{\text{time with novel}}{\text{total time}}$. This simple change in formula prevents artificial inflation of group-level effects or deviations from baseline and has been implemented by some researchers in the field (Cole, Ziadé, Simundic, & Mumby, 2020; Moreton et al., 2019). In general, when there is only one outcome of interest, analyzing that single outcome as a dependent variable is the simplest option to prevent a statistical violation. Another option for tasks with discrete trials is to analyze the data using mixed-effects binomial logistic regression (Cohen, 2002; Young, 2018).

However, choice paradigms with three or more outcomes of interest cannot be fixed with minor formulaic changes and require the use of analytic techniques that account for non-independence.

Choice Paradigms with Three or More Options

When choice paradigms have three or more options, the solution to interdependence is contingent on the number of outcomes of true interest. This problem can be illustrated by comparing two preclinical paradigms, the Morris Water Maze (MWM) and the Forced Swim Task (FST). The MWM is a measure of spatial memory (Morris, 1981) in which rodents are placed in a circular tank of opaque water and must locate (often using reference cues placed around the room) and swim to a platform just under the surface of the water. The maze can be divided into four equal quadrants, and a quadrant preference score is calculated. Traditionally, quadrant preference is defined as time spent in the target quadrant compared to time spent in other quadrants, and is used as a dependent variable in repeated measures ANOVA (Vorhees & Williams, 2006). This is problematic because it violates the assumption of independence; if percent time in one quadrant increases, percent time in the other quadrants inherently decreases. This is particularly important given the heavy reliance of the preclinical TBI field on MWM as a functional outcome. Some researchers have acknowledged this violation and use alternative types of analyses that do not rely on the GLM (Rogers, Churilov, Hannan, & Renoir, 2017). There is also another simple solution for this violation; rather than analyzing all four quadrants, researchers can quantify time spent in the target quadrant only. However, this only works because time spent in the target quadrant is generally the only outcome of interest.

In choice paradigms that have multiple outcomes of interest, such as the FST, the solution is less straightforward. The FST is a measure of depressive behavior during which rodents are placed in a cylindrical tank of water and become immobile after an initial period of swimming

and escape behaviors (Porsolt, Anton, Blavet, & Jalfre, 1978). It is common for these behaviors (swimming, floating, active escape) to be analyzed separately (i.e., separate *t*-test or ANOVA for each behavior) even though they are interdependent outcomes (e.g., Mezadri, Batista, Portes, Marino-Neto, & Lino-de-Oliveira, 2011). This is problematic not only due to the inflated number of tests, but because as any one behavior increases, the other behaviors inherently decrease. In other words, if there is an effect of a manipulation on time spent swimming, it is more likely there will be effects on floating and/or escaping. In a 2-choice paradigm, a solution is to analyze a single choice only. However, reducing FST outcomes to swimming only might fail to capture important information, given that different antidepressants can have differential outcomes on time spent swimming, climbing, and escaping (Slattery, Desrayaud, & Cryan, 2005). Thus, choice paradigms with distinct outcomes of interest present a complex problem. Another such task with multiple, distinct outcomes of interest is the Rodent Gambling Task (RGT) (Zeeb & Winstanley, 2013), which the Vonder Haar lab uses to assess chronic behavioral outcomes after TBI.

Rodent gambling task. The RGT is a choice paradigm with four interdependent outcomes of interest. It is a rat analogue of a neuropsychological assessment used to measure risky decision-making in clinical populations called the Iowa Gambling Task (Bechara, Damasio, Damasio, & Anderson, 1994). In the Iowa Gambling task, participants receive monetary gains and losses by choosing between four different decks of cards. In the RGT, rats can nosepoke in four different holes in an operant chamber. Each hole is associated with a different probability and magnitude of reinforcement (sucrose pellets) and punishment (timeout) and thus, a distinct distribution of risk and reward. As a result, there are four choice outcomes of interest: one optimal choice, two risky choices (with differential reinforcement/punishment), and

one suboptimal but non-risky choice. Most healthy control rats quickly learn to primarily choose the optimal hole (Zeeb & Winstanley, 2013), but TBI rats have persistent reductions in optimal choice (Shaver et al., 2019).

Statistical Approaches to Analyzing Choice Behavior

Traditionally, RGT data is analyzed by aggregating discrete trials into percent choice of each of the four options as the dependent variable for a repeated measures ANOVA or linear mixed-effects regression (LMER). Choice as a categorical variable (i.e., Choice 1, 2, 3, or 4, also referred to as P1, P2, P3, and P4 for the number of pellets delivered) is used as a predictor. If a manipulation (e.g., drug, injury) interacts with choice, dummy coding is used to relevel the choice variable and assess the effects of the manipulation on each choice option. A mixed model (also called multilevel, hierarchical, random effects) is typically used to analyze long-term outcomes on the RGT by incorporating both fixed effects and random effects. Fixed effects have a systematic effect on the outcome variable that is constant across individuals (e.g., group-level effects of TBI on symptoms). Random effects occur when levels of a variable are sampled from a larger population (e.g., individual patient effects, testing site effects). In repeated-measures RGT analyses, a mixed model can account for violations due to nested levels; LMER corrects the error by subtracting out variability that stems from individual-subject differences (Raudenbush & Bryk, 2001). Given the amount of between-subject variance seen on the RGT (Barrus, Hosking, Zeeb, Tremblay, & Winstanley, 2015), this is a major advantage. Another advantage of LMER is that it can effectively handle missing at random or missing completely at random data using restricted maximum likelihood estimation (Laird & Ware, 1982).

However, RGT data also violate the independence assumption because choice of any one option necessitates a shift in choice of other options. Violations that occur in other tasks (e.g.,

NOR, MWM) can be easily remedied because there is only one true outcome of interest (time spent with novel object and time in target quadrant, respectively). This solution does not apply to tasks with multiple outcomes of interest. For example, on the FST, it is important to measure time spent swimming, floating, and escaping; on the RGT, optimal, suboptimal, and risky choice are all outcomes of interest. Thus, LMER may not be the best approach to analyze data from these tasks, although it is important to note that RGT researchers *do not currently know* if this violation causes inaccuracies in practice.

One way to experimentally determine if a statistical test is (in)accurate is to simulate data. In the case of normality assumptions, some data simulations suggest that linear models can be applied to non-normal data in certain cases (Knief & Forstmeier, 2020). However, other papers show that linear models should *not* be applied to non-normal data. For example, one simulation study compared linear versus logistic regression for analyzing accuracy data constrained between 0 and 1 (as is percent choice on the RGT). Logistic regression performed on the raw correct/incorrect data outperformed linear models that treated accuracy as a continuous outcome (Dixon, 2008). Thus, the robustness of the GLM to violations is context dependent, and we must use simulations to empirically test the effects of the independence violation within the specific context of the RGT. For the current study, we tested the accuracy of LMER for analyzing RGT choice behavior. Notably, a multinomial logistic regression is a more appropriate technique for RGT data because it allows for analysis of raw data without violating the model assumptions. However, a multinomial logistic regression becomes computationally intensive when mixed effects are incorporated. Coercing choice data into a quasi-continuous variable and analyzing it using LMER or repeated-measures ANOVA is a much less computationally

intensive approach, but may be less accurate. The current study empirically tests the accuracy of this data coercion approach via simulation.

Data Simulations

Data simulations are an ideal method for testing the accuracy of different statistical tests. Monte Carlo methods are a common technique to simulate data through repeated sampling of random observations within a known probability distribution. This allows for the generation of data that mimic real-life processes (Kroese, Brereton, Taimre, & Botev, 2014). Monte Carlo methods are more prevalent in economics for determining the risk of different financial decisions; however, these methods have also been applied to psychological and biomedical sciences (e.g., Dixon, 2008; Meaney & Moineddin, 2014; Young, Cole, & Sutherland, 2012). One major advantage of Monte Carlo methods is that they can be used to evaluate the accuracy of different statistical tests because the researcher knows the "truth" of the data (i.e., whether data were sampled from equal or unequal distributions). For example, Monte Carlo simulations of MWM data demonstrated that both linear and non-linear mixed models identified real effects more accurately than ANOVA (Young, Clark, Goffus, & Hoane, 2009), and that a censored mixed model outperformed a linear mixed model (Young & Hoane, 2021).

Another major advantage of Monte Carlo simulations is that they can be used to simulate outcomes under a variety of different scenarios (Kroese et al., 2014). In behavioral neuroscience, these scenarios could be different sample sizes, effect sizes, or transformations applied to the data. Statistical tests have different advantages depending on sample size and effect size, particularly when analyzing non-normal data. For example, a large effect size ($\rho = 0.8$) was used to overcome small sample sizes ($n = 5$ and 10 per group) when analyzing non-normal data with Pearson and Spearman correlations (Bishara & Hittner, 2012). Sample size can also have a

substantial effect on the accuracy of mixed-effects models above and beyond power. For example, a large sample size (>50 subjects per group) can overcome the influence of rare events on a mixed-effects logistic regression model (Moineddin, Matheson, & Glazier, 2007). Thus, RGT datasets with various sample sizes and effect sizes must be simulated to determine if there are conditions under which LMER fails to overcome independence violations.

The Structure of Behavior

To simulate datasets that mimic real-life behavior, some knowledge about the structure of behavior is required. In the context of an operant task, such as the RGT, this involves information about what is driving choice. Although one might predict choice would be driven by optimal reinforcement (i.e., exclusive choice of the option with the highest reinforcement rate), this is not the case on the RGT. The field of behavior analysis suggests that choice is driven by both molar and molecular forces (Baum, 1989). The molar view asserts that the aggregate of reinforcement and punishment over many trials can be used to predict behavior. This theoretical orientation is further described by a behavior principle called the matching law, which states that the rate of responding on concurrent options is proportional to the rate of reinforcement provided by those options (Herrnstein, 1970). For example, a basketball player whose behavior is predicted by the matching law would spend more time shooting three-point shots because the rate of reinforcement (points gained per shot) for three-point shots is the highest. However, if the opposing team had a particularly strong defense against three-point shots, that player might attempt more two-point shots, thus matching behavior to the potential reinforcement offered by each outcome. In the context of the RGT, the molar view would be supported if rats chose each option proportionally according to reinforcement rates (i.e., chose P2 on about 44% of trials, P1 on about 31% of trials, and P3 on about 14% of trials and P4 on about 11% of trials).

In contrast, the molecular view states that immediate reinforcement and punishment on any single trial drives behavior (Shimp, 2020). This can be quantified in preclinical choice paradigms by examining "preference pulse," which calculates the degree of choice preference as a function of time from reinforcement (Davison & Baum, 2002) or, as is common in behavioral neuroscience, by calculating instances when a reinforced response is repeated on the next trial (i.e., win-stay) and when a punished response is changed on the next trial (i.e., lose-shift) (e.g., Stopper & Floresco, 2011). In the context of a basketball game, behavior would be consistent with the molecular view if a player decided between two-point and three-point shots based on the result of their previous shot. For example, if a player missed a three-point shot, they may shift (i.e., lose-shift) to a two-point shot next time they were in scoring range. In the context of the RGT, the molecular view would be supported if previous reinforcement and punishment exerted an outsized influence on the next trial (i.e., choice was always dictated by the outcome of the previous trial).

Structure of RGT Behavior.

To determine how to simulate realistic RGT behavior, in a published dataset (Vonder Haar, 2022b), we tested whether the data were reflective of either the molar or molecular view of quantitative prediction. To test the molar view, the matching law was fit to individual Sham and TBI subjects. Some Sham rats showed high sensitivity to overall reinforcement (i.e., steep linear increase in choice rate as reinforcement rate increased). However, some individual Sham subjects deviated from the matching law, showing an indifference to reinforcement rates or preference for options with a lower average reinforcement rate. TBI rats were more likely to deviate from matching both at the aggregate and individual subject level. Thus, although

aggregate choice generally resembled the shape of the matching law, other influences were unaccounted for by the molar view.

Next, we considered whether the molecular view accounted for RGT behavior by calculating the percentage of win-stay (e.g., if P2 choice was reinforced, the rat chose P2 on the next trial) and lose-shift trials (e.g., if P2 choice was punished, the rat chose a different option on the next trial). However, this theory of immediate reinforcement and punishment also did not account for behavior on the RGT. At the aggregate level, both Sham and TBI rats chose P2 more than chance regardless of whether the P2 option was previously reinforced or punished. Compared to Sham, TBI rats were less likely to stay regardless of a win or loss on the previous trial, potentially suggesting reduced molecular effects. This effect was moderated by choice option, such that TBI rats were more likely to stay on P1 but less likely to stay on P2. Overall, staying and shifting following wins and losses was inconsistent across individual subjects and did not explain deficits in TBI rats compared to Sham (Vonder Haar, 2022a; in press). Thus, neither the molar nor molecular view could account for behavior on the RGT, although molar forces did seem to outweigh molecular forces. It should also be noted that the aggregate data masked substantial individual subject variability in behavior which must be accounted for to understand choice. At the individual subject level, some intact rats actually preferred the riskier options. Thus, approaches that use behavioral theory to predict choice could not fully account for observed data on the RGT, and an alternative approach was needed.

Exploration-exploitation approach. Neither molar nor molecular prediction alone would likely be capable of simulating realistic RGT data. An alternative approach was to simply describe existing patterns in the data without those underlying assumptions. Choice data can be described using a theoretical approach from computational science called the exploration-

exploitation dilemma (Sutton, 1998). Exploitation can be defined as choice of the "best" option. Because rats on the RGT did not show consistent preference for the option with the highest overall reinforcement rate, we defined the "best" option as the most-preferred option based on our existing data. In an uncertain environment, identification of the most-preferred choice also requires exploration, or sampling different options. Visual inspection of our RGT data suggested that individual rats showed a different degree of preference for each option across repeated sessions and varied in the amount they exploited those preferences or explored among all the options. In human decision-making paradigms, this balance between exploitation and exploration has been quantified using an extension of Luce's choice axiom (Luce, 1959). Luce's choice axiom states that the probability of a response can be quantified by applying a weight (i.e., degree of saliency or preference) to that response. A set of weights may then be converted into probabilities using a mathematical function. A common application is the softmax transformation ($P_j = \frac{e^{\theta_j * weight_j}}{\sum_{i=1}^{k} \theta_i * weight_i}$), which takes in a list of values and uses an exponential function to transform those values into probabilities. To study the exploration-exploitation dilemma, cognitive psychologists have extended this softmax transformation of Luce's choice axiom to include a set of weights (or preferences) which is modified by a parameter representing the degree of exploitation versus exploration (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Namiki, Oyo, & Takahashi, 2015).

Although this softmax approach to the exploration-exploitation dilemma is rarely used in preclinical literature, it was used to explain choice behavior on an operant task by rats with frontal brain lesions (Dutech, Coutureau, & Marchand, 2011). Thus, a softmax function may be a good option to simulate data on other operant tasks, such as the RGT. Rather than using the molar and molecular theories of behavior to predict choice on the RGT, the softmax function can

be used to describe how individual subjects behave on the task. These descriptions can then be used to simulate data at the individual subject level. Applying the softmax function in this way would require knowledge of both the saliency/weight of each RGT choice and the degree of exploitation versus exploration for individual subjects. In the current study, we used existing data to quantify these softmax parameters to inform a simulation of RGT data.

Current Study

The current study consisted of two experiments with an overall goal of determining how interdependencies impacted RGT data analysis. The goal of Experiment 1 was to develop an effective method to simulate RGT data that mimicked the structure of observed behavior for Sham (i.e., un-injured) and frontal TBI rats. To generate these data, a descriptive approach using *k*-means clustering was used to determine preference weights and degree of exploitation/exploration across heterogenous subjects. These data were then passed into the softmax function to simulate trial-by-trial data. The goal of Experiment 2 was to identify conditions under which intercept-only LMER generated false positive and/or false negative results. To do this, simulations were repeated 1000 times at different sample sizes and magnitudes of TBI-induced deficits. It was hypothesized that intercept-only LMER would result in higher rates of false positives (i.e., greater than 5%) due to the interdependencies among choice outcomes on the RGT. This hypothesis was empirically tested using multiple Chi-Square tests where the outcome was the number of true (i.e., true positives or true negatives) versus false (i.e., false positives or false negatives) cases for each sample size and effect size.

Common Methods

Overview

RGT data was simulated (Experiment 1) and analyzed (Experiment 2) using R statistical software (https://www.r-project.org/). Simulations reflected actual rodent behavior on the task based on data we collected in the Vonder Haar lab. The methods (standardized across the field) for collecting RGT data are described below.

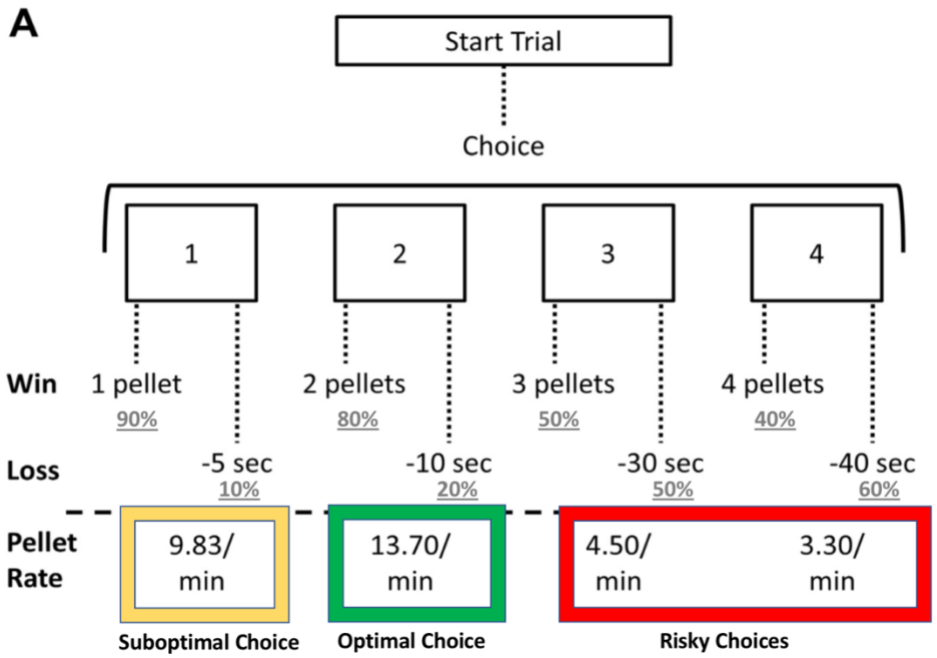
Rodent Gambling Task

The RGT was conducted in a standard operant chamber with a 5-hole array, but only four options (i.e., four holes) were presented in each trial. The one-pellet option (P1; non-risky but suboptimal) had a 90% probability of reinforcement and a 10% probability of a 5-s timeout from reinforcement. The two-pellet option (P2; optimal) had an 80% probability of reinforcement and a 20% probability of a 10-s timeout from reinforcement. The three-pellet option (P3; risky) had a 50% probability of reinforcement and a 50% probability of a 30-s timeout, and the four-pellet option (P4; risky) had a 40% probability of reinforcement and a 60% probability of a 40-s timeout. During the timeout, no responses were reinforced, and the light in the previously-chosen hole slowly flashed for the duration (1 Hz). A schematic of the task can be seen in Figure 1 (Shaver et al., 2019). The location of the P1, P2, P3, and P4 holes were counterbalanced across animals to account for potential side bias. All rats began with 7 sessions of a forced-choice procedure, which ensured each of the options were sampled equally. Then, rats progressed to the full task, where they were allowed to freely choose between the 4 choices for daily 30-min sessions with a maximum of 250 trials per session. The stimulus lights in the array were illuminated at the beginning of each trial. A response in any hole turned off the stimulus lights

and resulted in either reinforcement (sucrose pellets) or punishment (timeout from reinforcement) according to the probabilistic schedule described above.

Figure 1.

A Schematic of Reinforcement and Punishment Rates on the RGT.



A schematic of the Rodent Gambling Task (RGT). After initiating a trial, rats chose from any of the four holes. Each hole was associated with a different probability and magnitude of reinforcement and punishment (Shaver et al., 2019). As a result of varying reinforcement rates, the 1-pellet option (P1) was suboptimal, the 2-pellet option (P2) was optimal, and the 3- and 4-pellet options (P3, P4) were risky.

Methods: Experiment 1

Description and Design

The purpose of Experiment 1 was to simulate behavioral data for both Sham and TBI rats on the RGT. A compilation of RGT data (Vonder Haar, 2022b) was used to understand the structure of behavior on the task. This dataset, subsequently referred to as the control set, contained pre- and post-injury RGT data from 5 preclinical experiments with 151 adult male subjects (n = 71 for TBI; n = 80 for Sham). The control set contained trial-by-trial data for each subject across several weeks of sessions ranging from 2 to 12 weeks post-injury. For the current study, pre-injury sessions and data involving experimental manipulations (e.g., drugs) other than brain injury were excluded. This resulted in a total of 109 subjects (n = 58 for TBI; n = 51 for Sham). Only stable data from these subjects (i.e., data collected outside of the initial task-learning phase and acute injury phase) were considered.

To simulate trial-level data, the *rmultinom* function in R was used. This function takes in a list of probabilities and outputs multinomially distributed choices. To generate probabilities, weights (i.e., average rates of P1-4 choice) and an exploitation-exploration parameter (θ) were quantified from the control set, as per Luce's law of decision-making (see formula below). These weights and θ values were converted into probabilities using a softmax transformation and passed through the *rmultinom* function to generate discrete choices. However, these parameters could not be generated uniformly across subjects due to heterogeneity in individual-subject behavior. To account for considerable variability across subjects, behavioral phenotypes were extracted using *k*-means clustering (see below), and softmax parameters were defined separately for each cluster, resulting in simulated rats with distinct choice profiles that reflected the heterogeneity of the control set.

K-Means Clustering

The control set was used to identify unique choice phenotypes on the RGT for both Sham and TBI rats to be recapitulated in the simulation. Phenotypes were extracted using *k*-means clustering, an unsupervised learning algorithm that partitions data into *k* groups that cluster around a centroid, or cluster mean (Dwivedi, 2019). The number of clusters was determined using the elbow method, which involves fitting the data to a range for *k* (number of clusters) and plotting the error against *k.* The point of inflection in the plot determined the optimal number of clusters. Because the elbow method and supplemental techniques (e.g., average silhouette method) indicated a wide range of potential cluster numbers, additional considerations were used to determine cluster number as per our previous analysis of the control set (Vonder Haar, 2022a; in press). Specifically, a cluster number was only selected if the resulting clusters contained at least 5% of the total subjects. Each cluster was then considered a distinct behavioral phenotype and was referred to as a phenotype in text. The softmax function was fit to the control set data separately for each phenotype.

Softmax Function

A softmax function is a common machine learning transformation that takes in a list of values and returns probabilities. Modifications can be made to the softmax function to reflect the principles of the exploitation-exploration dilemma (Luce, 1959; Namiki et al., 2015). In the current study, probabilities of the four RGT choice options were calculated using the following softmax equation: $\text{Probability}_j = \frac{e^{\theta_j * weight_j}}{\sum_{i=1}^{4} \theta_i * weight_i}$, where the weight was a list of values representing the saliency or preference of each choice option and θ was the degree of exploitation versus exploration. Because each phenotype had different choice profiles and did not match the

true RGT reinforcement rates, average rates of choice by phenotype, rather than overall reinforcement rates, were used to calculate the weight parameter.

To generate plausible data, three types of parameters for each phenotype were calculated for both the weights and θ: (1) the population-level phenotype parameters (2) the between-subject variance, and (3) the within-subject variance. The population-level parameters were the average percent choice of P1-4 and average θ value for each phenotype calculated from the control set. Average choice was directly calculated from the control data (i.e., arithmetic mean for each phenotype), and the θ parameter was fit to individual subjects using the *nls* function in R, which determined the non-linear least-squares estimates of the parameters of a nonlinear model (in this case, the softmax function). To account for variance across subject and session, a between-subject and within-subject standard deviation (SD) were calculated for the weights and θ of each phenotype. Figure 2A shows the structure of the code where the population-level parameters and between-subject and within-subject variance were used, and Figure 2B shows more detailed pseudocode for using these parameters. For simplicity, functions were written outside the main body of the code to set the population-, subject-, and session-level parameters. These functions were then called within the code in an iterative fashion.

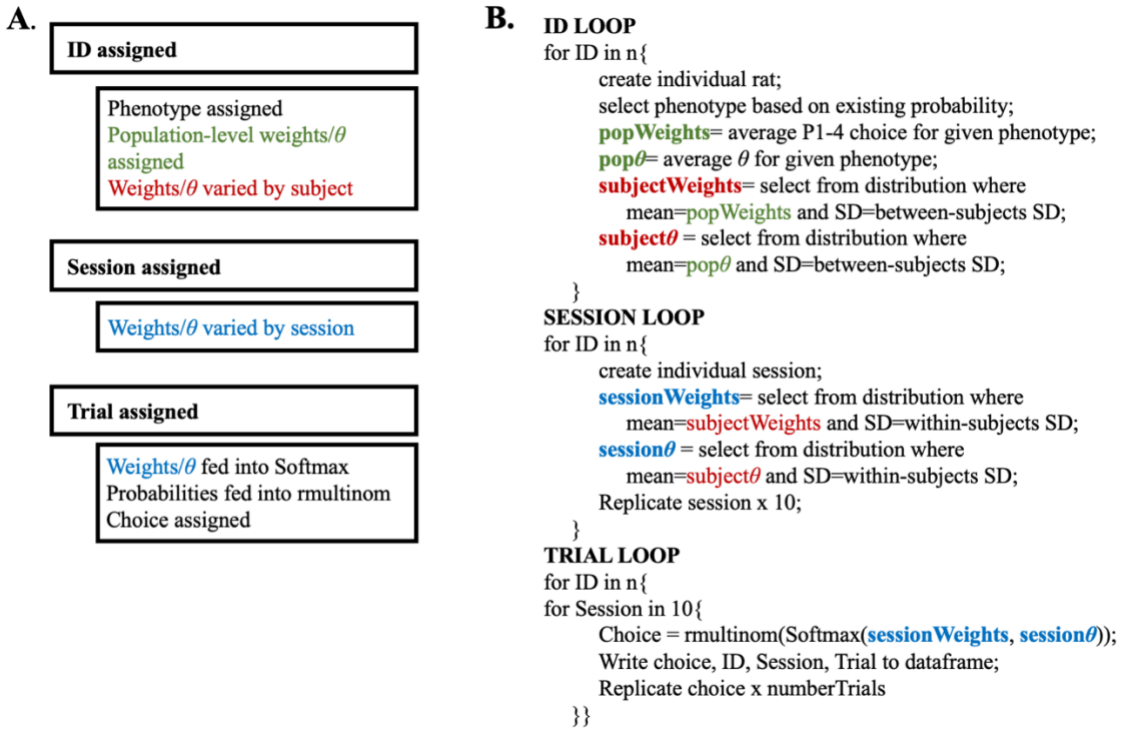
Iterative Simulation

Once the population-level parameters were calculated and varied by subject and session, the weights and θ were passed through the softmax function and converted into probabilities of each choice option on the RGT. Then, these probabilities were passed through the *rmultinom* function to generate discrete choices for each trial. To automate the process of varying weights and θ across subject and session, the data were simulated using a combination of the *replicate* function (which performs repeated evaluation of an expression) and nested for-loops (Figure

2A). The outermost loop created subjects, the middle loop created sessions, and the innermost loop created trials.

Figure 2.

Structure of R Syntax to Simulate Behavioral Data on the RGT.



Nested for-loop structure for the data simulation. Panel A shows the overall logical steps of the code. Subjects were assigned an ID and a phenotype. Then, population-level weights and θ (green text) were assigned to that subject, based on their phenotype. The weights and θ were varied for each subject (red text) and for each session (blue text). On any given trial, the weights and θ were passed through the softmax function, which generated probabilities that were fed into the *rmultinom* function to generate a discrete choice (either P1, P2, P3, or P4). This repeated for the assigned number of trials, and then a new session was generated until 10 sessions were complete. Panel B shows an example of pseudocode, where the population-level parameters of a phenotype were varied by subject and session and ultimately fed into the *rmultinom* function to generate trial-by-trial data. The softmax parameters for *n* subjects and 10 sessions were created first. Then, the replicate function was used to repeatedly pass those parameters through the *rmultinom* function for each trial. Color coding is consistent across panels, such that green text represents the population-level parameters for weights/θ, red text represents the subject-level weights/θ, and blue text represents the session-level weights/θ.

ID generation. Starting at the outermost subject loop, an ID number was selected. Then, that subject was assigned a phenotype based on the probability of that phenotype existing in the control set. The phenotype remained the same for any given subject at all levels of code. Once a phenotype was selected, a list of weights and a single θ parameter were sampled from a truncated normal distribution, where the mean was the population-level mean for that phenotype, and the SD was the between-subjects SD. The minimum and maximum of the truncated normal distribution were the minimum and maximum of the weights and θ observed in the control set. This resulted in distinct choice profiles for each phenotype, where each subject within a phenotype varied slightly from one another.

Session generation. Once a subject was assigned a phenotype, weights, and θ, the simulation progressed to the next loop. At this middle loop, 10 stable sessions of data were generated. To simulate variability across session, the weights and θ were varied slightly for each session, according to the within-subject SDs that were calculated from the control set. For each session, a new list of weights and a single θ parameter were sampled from a truncated normal distribution, where the mean was the weights/θ calculated in the subject loop and the SD was the within-subjects SD. This preserved the differences in choice profiles across phenotypes and subjects, while adding some variability across sessions.

Trial generation. Lastly, the simulation progressed to the trial-level loop that iterated through each subject and session. Trial number was sampled from truncated normal distributions for each phenotype, so that the number of trials reflected the data in the control set. This accounted for subtle differences in trial number across phenotypes (i.e., rats with riskier preferences incurred more timeouts and had fewer trials on average). Within the body of the loop, an RGT choice of P1, P2, P3, or P4 was selected for each discrete trial by passing the

weights and θ through the softmax function, which generated a probability of each choice. These probabilities were passed through the *rmultinom* function, which selected a discrete choice as a function of those probabilities. The choice for each trial, session, and subject was written to a data frame. This process was then repeated for the selected number of trials using the *replicate* function.

Data Processing and Visual Inspection.

Trial-level data was aggregated to a frequency count of each choice option per session and then converted to percent choice. Choice profiles were plotted against the control set and visually inspected to determine if reasonable data were generated. Data were inspected at the aggregate level (faceted by injury and phenotype) and at the individual subject level to ensure that the patterns in the control set were recapitulated in the simulation. Distributions of the within-subject and between-subject standard deviations were also plotted for P2 (optimal) choice. These plots were visually inspected by one primary rater and confirmed by two other raters. Methods were updated when there were visual discrepancies between the simulated data and the control set. For example, session-level variability was included in the simulation because initial simulations did not fully capture the variance in the control set.

Methods: Experiment 2

Description and Design

The purpose of Experiment 2 was to determine how rates of false positives and negatives using LMER changed across sample size and magnitude of TBI effect. To achieve this, 1000 datasets were simulated for each sample size and effect size as per similar designs (Burton, Altman, Royston, & Holder, 2006; Morris, White, & Crowther, 2019). Alpha (α) was set to 0.05 to determine whether a predictor had a significant effect.

Simulation Parameters

Sample size. Data were simulated for 4 sample sizes (n = 6, 10, 14, and 20) relevant to preclinical literature. Different sample sizes were generated by manipulating the number of subjects created in the R code. Preliminary simulations for the most extreme sample sizes (i.e., n = 6 and n = 20) were conducted first. Based on these results, two intermediary sample sizes were tested (i.e., n = 10 and n = 14), and additional sample sizes were not necessary.

Effect size. Effect size was less straightforward, given that we saw a shift in phenotypes rather than a net effect of TBI on each subject (Figure 3B). When considering the effect on P2 choice only, the TBI effect size in the control set was Cohen's f = 0.43. This effect size was recapitulated in the simulation by manipulating the probability that a subject belonged to a given phenotype. *K*-means clustering on the control set showed that TBI reduced the number of subjects that primarily selected the optimal P2 choice. Thus, the probability of belonging specifically to this high P2-preferring phenotype was used to generate various effect sizes. A standard TBI effect was generated by simulating TBI data with probabilities of phenotype prevalence that reflected the TBI data in the control set (i.e., a decrease in the high P2-preferring phenotype from approximately 60% to 20%). Then, an effect size of f = 0 was generated by

simulating TBI data with probabilities of phenotype prevalence that reflected the Sham data in the control set. Effect sizes above (f = 0.5) and below (f = 0.3) the observed TBI effect were generated by shifting the probability of the high P2-preferring phenotype and evenly distributing the difference across the other phenotypes, as observed in the control set (Figure 3B). However, this approach did not allow for the generation of exact effect sizes. To ensure that effect sizes were in a desired range, three datasets with n = 60 per injury condition (reflecting the size of the control set) were generated using an initial guess for the phenotype probability values. The size of the injury effect on P2 was then calculated using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) for each dataset. If the calculated effect size was in the desired range (f = 0.3-0.35, f = 0.4-0.45, and f = 0.5-0.55), for all three datasets those phenotype probability values were used for the 1000 datasets. These datasets were generated for 16 different conditions in total (4 sample sizes x 4 effect sizes).

Analysis of 1000 Datasets

Discrete trials were aggregated into percent choice of each RGT option as per Experiment 1, and an arcsine square-root transformation was applied to normalize data as in prior publications (Shaver et al., 2019). Then, LMER was conducted for each dataset using the *lme4* library in R (Bates, 2015). To determine the effects of interdependencies, two LMER models were evaluated. The first model (i.e., the interdependent model) tested the effects of TBI, Choice Option, and session on the transformed percent choice of each outcome. This model is subsequently referred to as the "full" LMER model, meaning that all four choices were analyzed simultaneously with subject as a random intercept (as opposed to a "full" random-effects structure). In this full LMER model, the TBI*Choice Option interaction was isolated to determine whether injury had a significant effect on choice. The second model (control LMER)

was subsetted to P2 choice only to determine the effect of TBI and session on P2 choice. This subsetted model served as a control because effect size was generated from the P2 variable and because it did not violate the independence assumption. The TBI main effect was isolated to determine whether injury had a significant effect on choice for the control model. For both models, the random effect was subject, with only the intercept varying across individual subjects. The *F*-statistic and *p*-value for these variables of interest were written to a dataframe for each analysis. Warnings and error messages (e.g., convergence failures) for each analysis were also written to a dataframe using the error-catching functions in the *purrr* library (Henry, 2020).

The two primary outcomes of interest were false positives and false negatives (Type I and Type II error respectively), which were calculated separately because they are independent. More specifically, when the simulated effect size was zero, there were two possible outcomes: TBI effect not expected/not observed (true negative) or TBI effect not expected/observed (false positive; Type I error). The frequency of false positives was visualized for each sample size, with an expected value of 50/1000 ($\alpha = 0.05$). It was predicted that false positives would exceed this rate for the full LMER model only. For the other effect sizes, the two possible outcomes were TBI effect expected/observed (true positive) and TBI effect expected/not observed (false negative; Type II error). The frequency of false negatives was visualized for each sample size and effect size. The expected values were determined via power analysis using G*Power (Faul et al., 2007). Results from the full LMER and control LMER were compared against the expected values using Chi-Square tests. These comparisons were performed separately for each sample size and effect size, and Bonferroni corrections were applied to adjust the *p*-values for multiple comparisons (Table 2).

Data Analysis on Single Datasets

To explore differences across multiple analytic techniques, a single dataset was randomly selected for each sample size and effect size (16 datasets total). Based on visual inspection, if a dataset was an outlier for the given effect size, a new set was selected at random. These datasets were analyzed using four approaches, and test statistics were reported in Tables 3-6. The first and second approaches were the full LMER model (Table 3) and the control LMER model (Table 4). The third approach was a generalized linear mixed model with a logit link function, subsequentially referred to as a binomial logistic regression. Choice data was recoded into a dichotomous variable where the two levels were P2 versus all other options and treated as a proportion (P2 choice/total choice). The fixed effect in the model was injury, the random effect was subject (intercept only), and the outcome was choice (P2 vs. others). The *glmer* function in R was used to perform a weighted binomial logistic regression, and the resulting test statistics (log odds) and *p*-values were reported in Table 5.

The fourth analysis was a multinomial logistic regression with a Bayesian approach using the *brms* package in R (Bürkner, 2018). The fixed effect in the model was injury, the random effect was subject (intercept only), and the outcome was choice (categorical variable with four levels), which was also analyzed at the proportion level. The Bayesian models were generated with the default priors from the *brms* package. The range of parameters composing the prior and posterior distributions was selected using Markov chain Monte Carlo sampling with four chains. Rather than converging on single regression parameter estimates and their uncertainty (standard errors) as per standard frequentist statistics, a range of values (called a posterior probability density distribution) most likely to contain the population-level regression parameters was generated. The *emmeans* library (Lenth, 2021) was used to calculate the most likely estimate (log odds) for the effect of injury on P2 choice and the 95% credible interval (Table 6). The credible

interval differs slightly from a confidence interval in null hypothesis testing. There is a 95% chance that the true population statistic falls within the range of values in a 95% credible interval (Kruschke, 2014; Young, 2019). Typically, the credible interval would not be used to make a categorical decision of significant versus non-significant effects. However, to compare the results of the Bayesian analysis with the binomial and linear models, an effect was considered significant if the credible interval did not contain zero.

Results: Experiment 1

K-Means Clustering

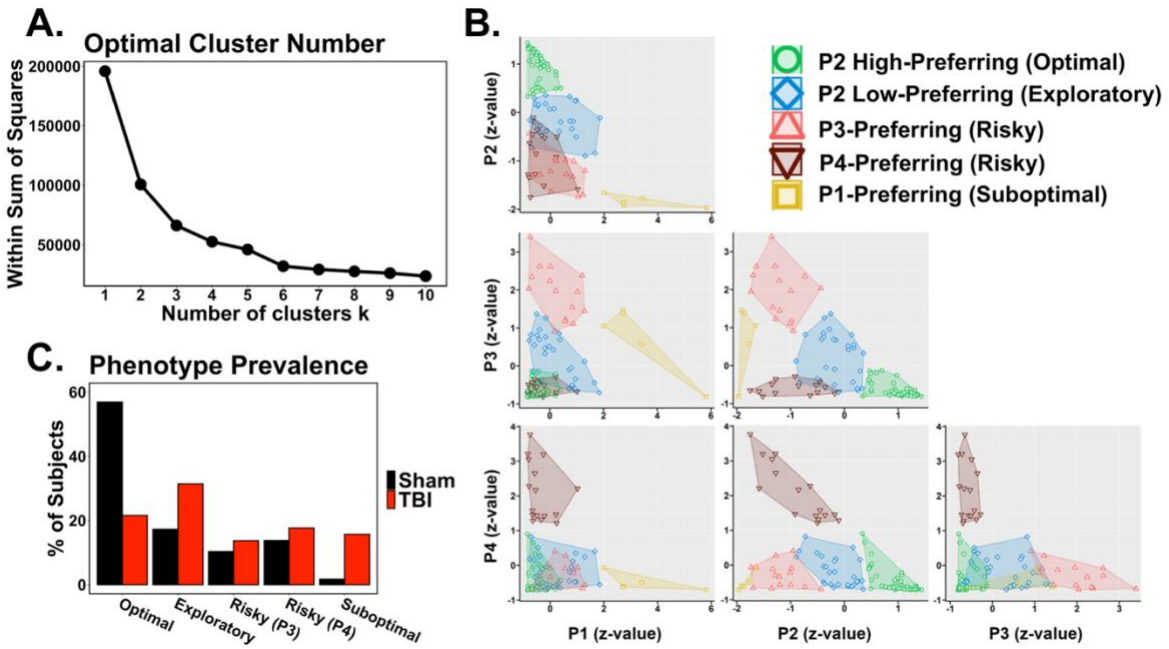
The goal of Experiment 1 was to simulate RGT data, which required *k*-means clustering to extract phenotypes from the control set. An elbow plot of the within sum of squares for a range of cluster numbers indicated that the optimal cluster number (*k*) was between 2 and 6 clusters (Figure 3A-B). The number of subjects within each cluster was then calculated for this range of cluster numbers. To maximize the variance explained by clusters while also preventing overfitting, the largest cluster number *k* that resulted in at least 5% of subjects within each cluster was selected. When TBI and Sham data were clustered together, $k = 6$ resulted in some clusters that contained less than 5% of subjects. However, $k = 4$ resulted in imprecise group-level fits (i.e., there were visually apparent differences between Sham and TBI rats within the same cluster). Thus, five clusters ($k = 5$) were selected because it maximized variance explained in the data without overfitting and resulted in choice profiles that were consistent across Sham and TBI rats within a cluster.

Based on visualizations of RGT choice profiles, each cluster was referred to as a phenotype and assigned a unique descriptor (Figure 3C). The phenotypes were (1) high P2-preferring/optimal (2), low P2-preferring/exploratory, (3) P3-preferring/risky, (4) P4-preferring/risky, and (5) P1-preferring/suboptimal. Sham rats primarily belonged to the optimal phenotype. After TBI, the prevalence of the optimal phenotype was reduced from approximately 60% to 20% with roughly even redistributions to each of the other four phenotypes (Figure 3D). These results were then used to inform the simulation parameters for Experiment 1 and 2. Because TBI caused a shift in the distribution of individual choice profiles, rather than an overall

net reduction in optimal choice for each subject, the "effect size" in Experiment 2 was manipulated by adjusting the distribution of choice profiles.

Figure 3

K-Means Clustering on the Control Set of RGT Data.



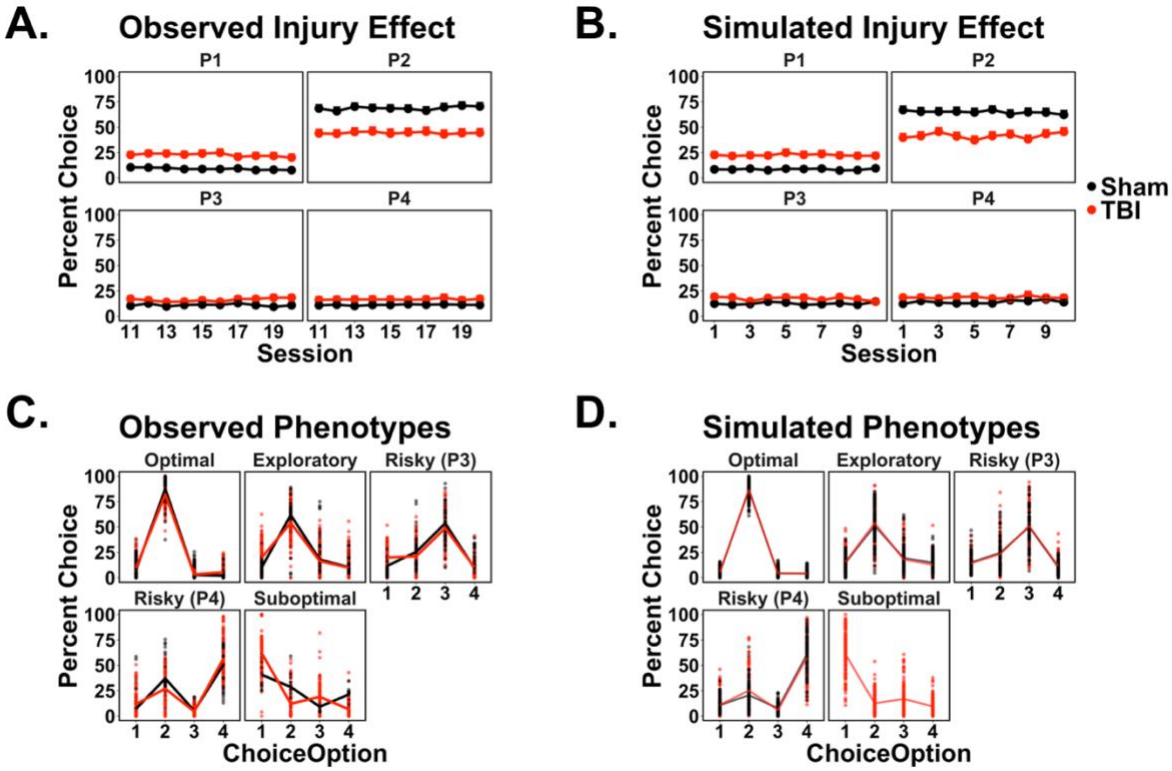
K-means clustering results for the control set. Panel A shows the elbow plot of the total within sum of squares as a function of cluster number. To create distinct clusters without overfitting that were consistent across Sham and TBI rats, the value *k* (number of clusters) was set at 5. Panel B shows the distinct choice profiles of each of the five phenotypes. The x- and y-axes show the *z*-scores for the average choice of P1, P2, P3, and P4 within a phenotype to distinguish between the P2 high-preferring (optimal; shown in green circles), P2 low-preferring (exploratory; shown in blue diamonds), P3-preferring (risky; shown in red triangles), P4-preferring (risky; shown in burgundy inverted triangles), and P1-preferring (suboptimal; shown in yellow squares) phenotypes. Panel C shows the prevalence of each phenotype for Sham (black) versus TBI (red) rats in the control set. The optimal phenotype (P2 high-preferring) is most prevalent for Sham rats, but decreases in prevalence for TBI rats.

Preliminary Simulations

Preliminary simulations were conducted using softmax parameters that varied by phenotype, subject, and session. Exemplar datasets were generated for Sham and TBI rats (n = 60 per group). Simulated data closely approximated the observed data when plotted by injury (Figure 4A-B) and by phenotype (Figure 4C-D). The simulation captured effects that have been replicated in our observed data: TBI increased suboptimal choice, decreased optimal choice, and had inconsistent effects on risky choice. A minor discrepancy at the phenotype level was that simulated Sham and TBI rats appeared more similar (i.e., lines and points were more overlapping in Figure 4D) compared to the observed data (Figure 4C) in the control set. This discrepancy was expected because the same parameters were used to generate Sham and TBI rats (with the only difference being the probability of belonging to a given phenotype), which did not affect core questions surrounding power and false positive rates.

Figure 4.

Preliminary Simulations of RGT Data



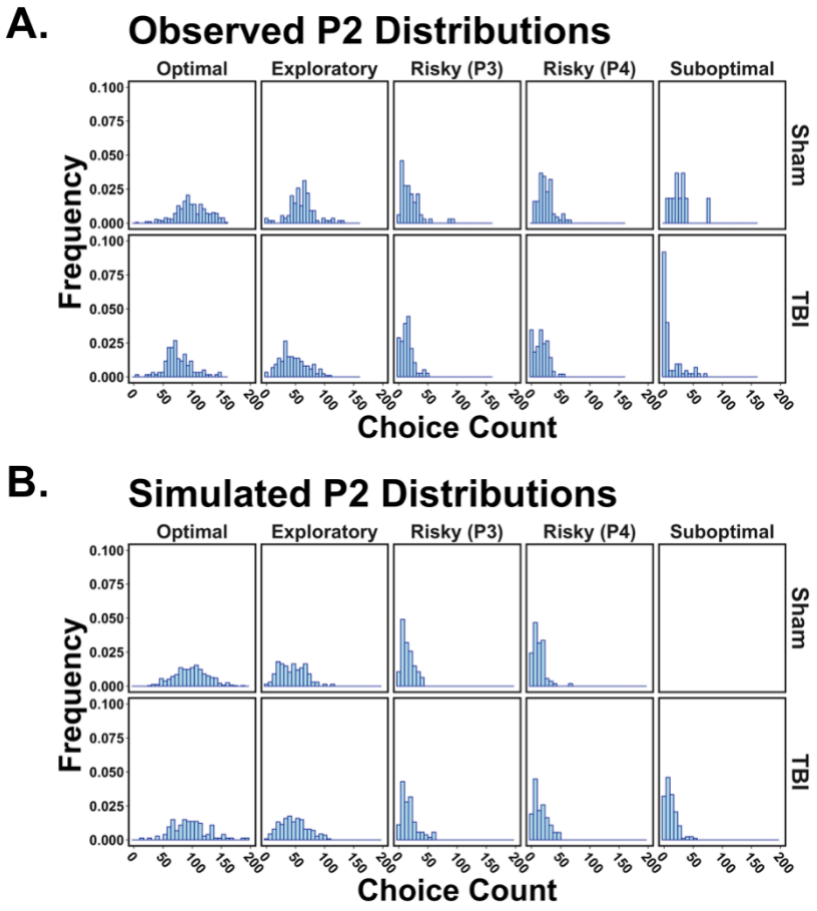
Simulated RGT data compared against observed RGT data. Panel A shows observed choice of each option (P1, P2, P3, and P4) for Sham (black) versus TBI (red) rats in the control set. Panel B shows simulated choice of each option for Sham versus TBI rats. Data shown in Panels A and B are mean+SEM. Panels C and D show individual (points) and average (lines) choice of each option faceted by phenotype for observed data in the control set and simulated data, respectively.

To further ensure that simulated data were consistent with observed data, visual inspection was performed for individual subject data and for the distributions of P2 choice. Visual inspection showed strong concordance between simulated and observed data at the individual subject level. Histograms were used to visualize P2 distributions and showed a close concordance between simulated and observed data for both Sham and TBI (Figure 5 A-B). Additional histograms were generated to view distributions of the within-subject and between-subject standard deviation for P2 choice, as the variance heavily impacts subsequent statistical

analyses. These histograms also showed a reasonable concordance with some minor expected discrepancies between simulated and observed data for both Sham and TBI (Figure 6 A-D). All three raters agreed that the simulated data was an accurate reflection of the control set data.

Figure 5.

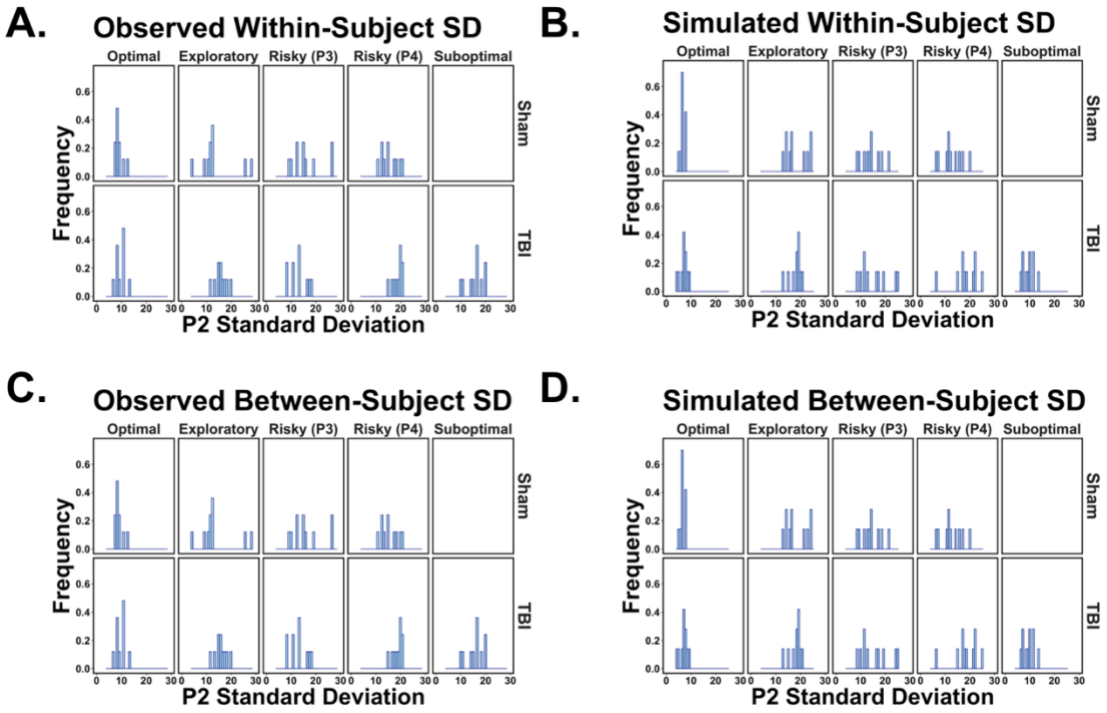
Distributions of P2 Choice in Preliminary Simulation



Distributions of P2 choice in observed (Panel A) and simulated (Panel B) RGT data across the five phenotypes. The x-axis shows the number of trials within a session where P2 was selected, and the y-axis shows the frequency of each specific choice count ranging from 0-10%. The upper five panels contain the distributions for Sham rats, and the lower panels show TBI rats. In this single simulated dataset (Panel B), no suboptimal rats were generated for the Sham group due to the low prevalence of this phenotype in the control set (one subject only). However, some suboptimal sham rats were simulated separately to ensure that choice distributions reflected the control set data.

Figure 6.

Distributions of P2 Standard Deviations in Preliminary Simulation



Distributions of P2 standard deviations. The observed (Panel A) and simulated (Panel B) within-subject standard deviations and observed (Panel C) and simulated (Panel D) between-subject standard deviations for P2 choice are shown for each phenotype. The upper five panels contain the distributions for Sham rats, and the lower panels show TBI rats. The SDs could not be calculated for the observed suboptimal Sham data (Panels A and C) because there was only one subject. The SDs were not shown for suboptimal Sham rats in the simulation (Panels B and D) because no rats were assigned this phenotype in the single simulation by chance.

Results: Experiment 2

The goal of Experiment 2 was to determine how rates of false positives/negatives varied across sample sizes and effect sizes when using LMER to analyze RGT data. Two LMER approaches were tested on 1000 datasets per condition. The Chi-Square results comparing each approach to the expected values are provided in Table 2. The full LMER was significantly different than expected values across all sample sizes and effect sizes (Table 2), with a high propensity for "hits" (false positives and true positives) (Figure 7). The subsetted LMER was not significantly different than expected results for all sample and effect sizes (Table 2). Because of the substantial false positive rate observed with the full LMER, a subset of datasets (1 per condition; 16 total) was randomly selected to analyze with additional categorical methods. The two LMER approaches were also applied to these single datasets to serve as a comparison. Each model is described in further detail below.

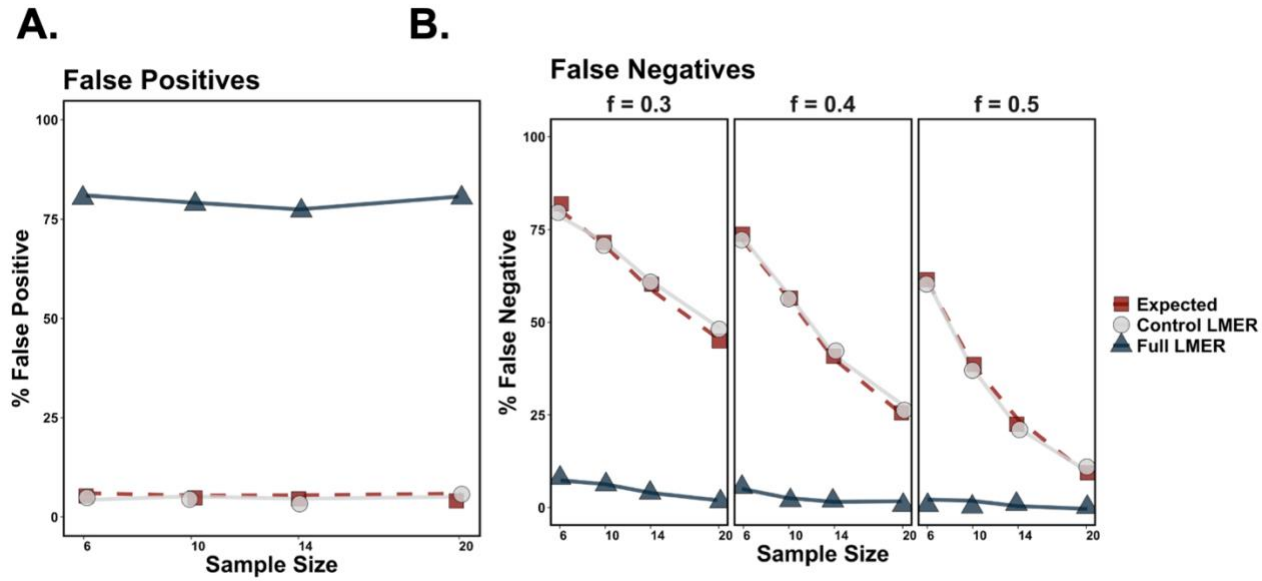
Table 2. Chi Square Tests Comparing the Full and Control LMER Against Expected Results.

		Full vs. Expected			Control vs. Expected		
Effect	Sample	Chi-square	Unadjusted *p*	Corrected *p*	Chi-square	Unadjusted *p*	Corrected *p*
$f = 0.0$	$n = 6$	1161.797	<0.00001	<.001	1.163	0.281	>0.99
	$n = 10$	1102.116	<0.00001	<.001	0.041	0.839	>0.99
	$n = 14$	1068.942	<0.00001	<.001	1.163	0.281	>0.99
	$n = 20$	1170.024	<0.00001	<.001	0.162	0.697	>0.99
$f = 0.3$	$n = 6$	1072.477	<0.00001	<.001	1.016	0.313	>0.99
	$n = 10$	901.774	<0.00001	<.001	0.039	0.843	>0.99
	$n = 14$	742.402	<0.00001	<.001	0.354	0.552	>0.99
	$n = 20$	523.777	<0.00001	<.001	1.257	0.262	>0.99
$f = 0.4$	$n = 6$	971.837	<0.00001	<.001	1.107	0.293	>0.99
	$n = 10$	691.553	<0.00001	<.001	0.073	0.787	>0.99
	$n = 14$	482.218	<0.00001	<.001	0.074	0.785	>0.99
	$n = 20$	260.611	<0.00001	<.001	1.040	0.308	>0.99
$f = 0.5$	$n = 6$	744.043	<0.00001	<.001	0.539	0.463	>0.99
	$n = 10$	448.039	<0.00001	<.001	0.415	0.520	>0.99
	$n = 14$	256.662	<0.00001	<.001	0.741	0.389	>0.99
	$n = 20$	102.201	<0.00001	<.001	0.064	0.801	>0.99

Note: For the effect size $f = 0.0$, the outcome is false positive versus true negative. For all other effect sizes, the outcome is false negative versus true positive. Both unadjusted and Bonferroni-corrected *p*-values (original *p*-value multiplied by 4) are provided. The full LMER was significantly different than expected values across all conditions. The control model was not significantly different than the expected values across all conditions.

Figure 7.

LMER Analyses of Simulated RGT Data



Rate of false positives (Panel A) and false negatives at three effect sizes (Panel B). Expected values (red dashed line with square points) were compared to the subsetted LMER results (control LMER; gray solid line with circle points) and full LMER results (blue solid line with triangle points). For false positives, the control model was consistent with the expected error rate of 5% at all sample sizes. The full LMER model had a false positive rate exceeding 75% at all sample sizes. For false negatives, the control model was consistent with the expected power curve generated in G*Power, and false negatives decreased as sample size and effect size increased. The full LMER model had low false negative rates across all sample sizes and effect sizes. A jitter (width=0.15, height=1.0) was applied to the lines and points in both panels due to the overlap between the expected values and control model.

Full LMER model

The full LMER model was defined as Choice~Option*Injury*Session+(1|Subject). The main variable of interest was a significant Option*Injury interaction. At an effect size of zero, there were over 750 significant Option*Injury interactions (compared to the expected value of 50) (Figure 7A). Across all other effect sizes, false negatives with the full LMER model were substantially lower than the expected values from G*Power (Figure 7B). The difference between the full LMER outcome and expected outcome was statistically significant across every sample

size and effect size (Table 2; p's < 0.05). None of the models failed to converge, but all models had a singular fit warning. The beta coefficients, standard error, degrees of freedom, *t*-value, and *p*-value were provided for a single full LMER model for each sample size and effect size (Table 3).

Table 3. Single Dataset Evaluations: Full LMER Model (Injury Effect on P2)

Effect	Sample	Estimate	Error	df	*t*	*p*
f = 0.0	n = 6	-0.044	0.119	464	-0.367	0.713
	n = 10	-0.040	0.091	784	-0.441	0.659
	n = 14	0.243	0.073	1104	3.319	<0.001
	n = 20	0.293	0.065	1584	4.523	<0.001
f = 0.3	n = 6	-0.493	0.129	464	-3.812	<0.001
	n = 10	-0.598	0.109	784	-5.506	<0.001
	n = 14	-0.778	0.094	1104	-8.249	<0.001
	n = 20	-0.308	0.0769	1584	-4.007	<0.001
f = 0.4	n = 6	-0.493	0.129	464	-3.812	<0.001
	n = 10	-0.800	0.110	784	-7.246	<0.001
	n = 14	-0.861	0.094	1104	-9.169	<0.001
	n = 20	-0.788	0.083	1584	-9.524	<0.001
f = 0.5	n = 6	-1.170	0.138	464	-8.471	<0.001
	n = 10	-1.018	0.111	784	-9.200	<0.001
	n = 14	-1.030	0.096	1104	-10.737	<0.001
	n = 20	-0.664	0.0806	1584	-8.236	<0.001

Note: The test statistics and *p*-values for the full linear mixed-effects regression (LMER) model.

Control LMER Model

The control LMER model (subsetted to P2 choice only) was defined as P2Choice~Injury*Session+(1|Subject). The outcome of interest was a significant main effect of injury. At an effect size of zero, false positives closely mapped onto the expected value of 50 (Figure 7A). At all other effect sizes, false negatives were consistent with expected values from G*Power (Figure 7B). There were no statistically significant differences between the control LMER results and expected results across any sample sizes or effect sizes (Table 2; p's > 0.05). None of the models failed to converge or produced any warnings or messages. The beta

coefficients, standard error, degrees of freedom, *t*-value, and *p*-value are provided for a single control LMER model for each sample size and effect size (Table 4).

Table 4. Single Dataset Evaluations: Control LMER Model (Injury Effect on P2)

Effect	Sample	Estimate	Error	df	*t*	*p*
f = 0.0	*n* = 6	0.011	0.580	12.176	0.020	0.985
	n = 10	-0.147	0.437	21.751	-0.336	0.740
	n = 14	0.243	0.373	28.573	0.650	0.521
	n = 20	0.292	0.304	43.654	0.963	0.341
f = 0.3	*n* = 6	-0.470	0.546	13.153	-0.862	0.404
	n = 10	-0.589	0.429	20.471	-1.372	0.185
	n = 14	-0.704	0.349	29.378	-2.015	0.053
	n = 20	-0.245	0.307	42.794	-0.799	0.428
f = 0.4	*n* = 6	-0.470	0.546	13.153	-0.862	0.404
	n = 10	-0.707	0.409	21.196	-1.726	0.099
	n = 14	-0.786	0.342	29.564	-2.297	0.029
	n = 20	-0.749	0.290	42.861	-2.582	0.013
f = 0.5	*n* = 6	-0.794	0.449	15.522	-1.768	0.097
	n = 10	-0.927	0.388	22.095	-2.386	0.026
	n = 14	-0.970	0.324	31.158	-2.990	0.005
	n = 20	-0.589	0.293	43.405	-2.011	0.051

Note: The test statistics and *p*-values for the control linear mixed-effects regression (LMER) model.

Binomial Logistic Regression Model

The binomial logistic model was performed to determine the effect of injury on P2 choice versus all other choices for a single dataset at each sample size and effect size. The estimates (log odds), errors, *z*-test, and *p*-values are provided in Table 5. There were no false positives, but there were some false negatives at effect sizes of *f* = 0.3 and *f* = 0.4 only. This model was in strong agreement (87.5% concordance) with the Bayesian multinomial model.

Table 5. Single Dataset Evaluations: Binomial Logistic Regression (Injury Effect on P2)

Effect	Sample	Estimate	Error	*z*	*p*
f = 0.0	*n* = 6	0.007	0.681	0.011	0.991
	n = 10	-0.081	0.511	-0.159	0.874
	n = 14	0.476	0.537	0.886	0.375
	n = 20	0.507	0.427	1.186	0.235
f = 0.3	*n* = 6	-0.642	0.627	-1.025	0.305
	n = 10	-0.989	0.626	-1.580	0.114
	n = 14	-1.187	0.555	-2.140	0.032
	n = 20	-0.534	0.458	-1.165	0.244
f = 0.4	*n* = 6	-0.642	0.627	-1.025	0.305
	n = 10	-1.178	0.554	-2.124	0.034
	n = 14	-1.323	0.535	-2.479	0.013
	n = 20	-1.242	0.481	-2.582	0.009
f = 0.5	*n* = 6	-1.382	0.510	-2.711	0.007
	n = 10	-1.494	0.568	-2.631	0.009
	n = 14	-1.567	0.508	-3.084	0.002
	n = 20	-1.039	0.442	-2.353	0.019

Note: The estimates, errors, z-values, and p-values for the binomial logistic regression using the *glmer* function with a logit link in R.

Bayesian Multinomial Logistic Regression Model

The multinomial model was performed to determine the effect of injury on choice for a single dataset at each sample size and effect size. All Rhat values were less than 1.2, showing consistent convergence among the four chains. This was confirmed by visual inspection of the Markov chain traceplots. The estimates (log odds) and credible intervals are provided in Table 6. There were no false positives at any sample size. However, there were some false negatives at effect sizes of *f* = 0.3 and *f* = 0.4, There were no false negatives at *f* = 0.5. These results were compared against the full LMER, control LMER (subsetted to P2), and binomial logistic regression (Table 7). The full LMER model was in 50% concordance with the Bayesian model, whereas the control LMER and binomial model were in 87.5% concordance (Figure 8).

Table 6. Single Dataset Evaluations: Multinomial Logistic Regression (Injury Effect on P2)

Effect	Sample	Estimate	Lower	Upper	"Significant"
f = 0.0	*n* = 6	0.122	-1.550	1.720	No
	n = 10	0.099	-1.020	1.210	No
	n = 14	-0.473	-1.540	0.500	No
	n = 20	-0.379	-1.060	0.239	No
f = 0.3	*n* = 6	0.323	-1.070	1.660	No
	n = 10	1.170	-0.380	2.710	No
	n = 14	0.827	-0.268	1.870	No
	n = 20	0.597	-0.201	1.430	No
f = 0.4	*n* = 6	0.337	-1.000	1.680	No
	n = 10	1.080	-0.180	2.280	No
	n = 14	1.290	0.070	2.490	Yes
	n = 20	1.510	0.627	2.500	Yes
f = 0.5	*n* = 6	1.390	0.027	2.820	Yes
	n = 10	1.880	0.509	3.200	Yes
	n = 14	1.440	0.285	2.550	Yes
	n = 20	1.070	0.316	1.860	Yes

Note: The estimates and lower and upper confidence interval for the multinomial logistic regression analyses. The "Significant" column is marked as "no" if the confidence interval contained zero and marked "yes" if the interval did not contain zero. The latter was classified as a significant effect to allow for direct comparison with the other models.

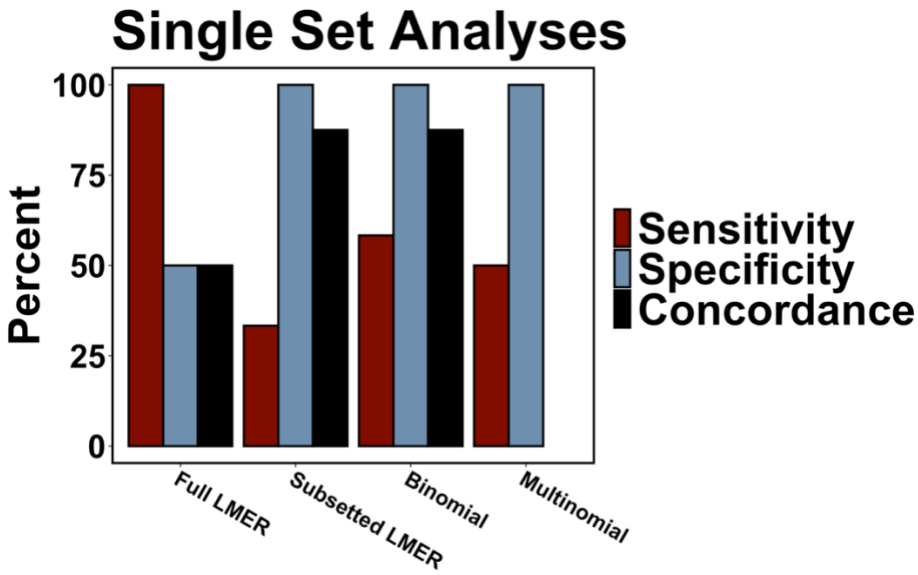
Table 7. Single Dataset Evaluations: Comparisons Across Techniques

Effect	Sample	Full LMER	Control LMER	Binomial	Multinomial
$f = 0.0$	$n = 6$	TN	TN	TN	TN
	$n = 10$	TN	TN	TN	TN
	$n = 14$	**FP**	TN	TN	TN
	$n = 20$	**FP**	TN	TN	TN
$f = 0.3$	$n = 6$	**TP**	FN	FN	FN
	$n = 10$	**TP**	FN	FN	FN
	$n = 14$	**TP**	FN	**TP**	FN
	$n = 20$	**TP**	FN	FN	FN
$f = 0.4$	$n = 6$	**TP**	FN	FN	FN
	$n = 10$	**TP**	FN	**TP**	FN
	$n = 14$	TP	TP	TP	TP
	$n = 20$	TP	TP	TP	TP
$f = 0.5$	$n = 6$	TP	**FN**	TP	TP
	$n = 10$	TP	TP	TP	TP
	$n = 14$	TP	TP	TP	TP
	$n = 20$	TP	**FN**	TP	TP

Note: The abbreviations in the table specify whether a true negative (TN), false positive (FP), true positive (TP), or false negative (FN) occurred for each sample size and effect size across different analytic techniques. Results that are *inconsistent* with the multinomial logistic regression are bolded.

Figure 8.

Single Dataset Analyses: Accuracy and Concordance with Multinomial Logistic Regression



Results of the single dataset analyses for the four analytic models: full LMER, control LMER, binomial logistic regression, and Bayesian multinomial logistic regression. The three outcomes shown are sensitivity (percentage of true positives; red), specificity (percentage of true negatives; blue) and concordance with the multinomial logistic regression (black). Although the full LMER had the highest sensitivity, it had the lowest specificity and was least concordant with multinomial logistic regression. The control LMER had the lowest sensitivity, but was more consistent with the multinomial model. The binomial logistic model had the best balance in sensitivity, specificity, and concordance.

Discussion

The goal of the current study was to empirically determine if simultaneously analyzing multiple choice outcomes on the RGT using a linear model was problematic and thus may implicate broader problems with analysis of choice behavior. The results demonstrated that a linear model using choice as a predictor and random subject intercepts was not suitable for the conditions tested. This 4-choice linear model exceeded 75% false positives for all sample sizes (Figure 7A). The comparison of this full LMER to the control LMER (subsetted to P2 choice only), which had approximately 5% false positives, provides further evidence that the independence violation artificially inflates effects to a very meaningful degree when using certain linear models to analyze RGT choice. Although the control model had an acceptable rate of false positives, it only accounted for one outcome. Because there are four outcomes of interest on the RGT (one suboptimal, one optimal, and two risky), the control LMER would need to be repeated four times, which would inflate the false positive rate to about 18.5%. To account for this increased family-wise error, corrections would be required, which is problematic due to the lower power of the control model. The control LMER was underpowered to detect the typical TBI effect size ($f = 0.4$) even at 20 subjects per group, a sample size much higher than the preclinical norm. In place of the full LMER model, two potential alternative approaches are (1) improving the accuracy of the control LMER and (2) using categorical analyses that do not assume independent outcomes.

Alternative Approaches

Improving LMER Accuracy

It is possible that a linear mixed-model can still be used to analyze RGT data. First, other parameters not tested in the current study could potentially increase the power of the control

model. These may be of interest to explore due to the heavy reliance of the behavior analysis field on linear models. It has been known for several decades that behavioral outcomes are often best described by non-linear models (Meddings, Scott, & Fick, 1989). Nonetheless, most behavior analysts have still not adopted non-linear approaches and continue to transform data into a quasi-continuous structure (e.g., aggregating trials into a percent choice). Thus, it may be useful to explore options that increase the power of the control LMER, such as additional sessions. Only 10 sessions of data were simulated in the current study, but we often conduct behavioral testing for over 50 sessions post injury. It is possible that increasing session number might increase the power of the control LMER, and thus decrease the false negative rate.

Another option for improving LMER accuracy is to change the random effects structure. Additionally, the current study used a random intercept-only model for both LMERs, which likely leaves some unexplained variance due to the exclusion of random slopes (Heisig & Schaeffer, 2018). Session as a random slope would likely have little influence on the results because no systematic effect of session was simulated. This is further supported by the fact that the control analysis (which was also an intercept-only model) mapped onto expected values for both false positives and false negatives. However, adding choice preference as a random slope (i.e., allowing the difference between P2 versus the other options to vary by subject) might attenuate the elevated false positive rate of the full intercept-only LMER. This is particularly important given that mixed models with a maximal random effects structure are more generalizable than intercept-only models (Barr, Levy, Scheepers, & Tily, 2013), although caution must be exercised to prevent convergence issues and uninterpretable models when expanding the random-effects structure (Bates, Kliegl, Vasishth, & Baayen, 2015). Nonetheless, it may be beneficial to explore how LMER accuracy for RGT data is affected by random slopes.

Categorical Approaches

However, there may not be any conditions that allow for the use of linear models to analyze RGT data. Thus, it is also important to explore categorical analyses, which are a truer fit to the raw structure of RGT data. An ordinal logistic regression theoretically fits the data given that the reinforcement rates on the RGT dictate an ordinal structure ranging from highest to lowest average reinforcement rate. However, the control set data, and particularly the phenotyping, demonstrated that choice does not always reflect an ordinal data structure for intact rats (Vonder Haar, 2022a; in press). Thus, a multinomial logistic regression is likely the truest fit to the data, as the outcome variable is categorical with more than two levels. A drawback is that this type of analysis is unfamiliar to most behavioral researchers. It is also computationally intensive to incorporate mixed effects into a multinomial logistic regression and was accomplished in the current study by using a Bayesian approach with the *brms* package in R. Some preclinical researchers might be reluctant to learn these types of Bayesian techniques, but biostatisticians might be engaged by the complexities of choice analysis and eager to collaborate. Furthermore, pilot analyses from the current study suggest that a mixed-effects *binomial* logistic regression may closely approximate the findings of a similar multinomial analysis (Figure 8). A Begg and Gray approach (Begg & Gray, 1984) could be used to compare P2 against all other choices, P1 against all other choices, etc. Ideally, this repeated pairwise approach might closely approximate a mixed-effects multinomial analysis without the computational intensity. However, this question remains unanswered because multinomial and binomial logistic regression were only compared for a single dataset per sample size and effect size in the current study.

To empirically determine if binomial logistic regression is suitable for RGT data analysis, a full examination of its accuracy with 1000 datasets will be required. If the binomial model has

a reasonable rate of false positives, it may be the most appealing approach. First, it may outperform the control LMER because it fits the categorical structure of RGT outcomes, and prior simulations have demonstrated that binomial logistic regression generally outperforms the percent choice approach for data bound between zero and one (Dixon, 2008). It is also preferable to the multinomial logistic regression because it is less computationally intensive and more familiar to behavioral researchers. However, a drawback of binomial logistic regression is that it requires multiple comparisons (i.e., P2 vs. other options, P1 vs. other options, etc.). Another drawback is that it is also outside of standard practice for behavioral researchers. Regardless, the full LMER is not a suitable approach for analyzing RGT data, and published data may contain inaccurate results.

Published RGT Literature

Based on the findings of the current study, a review of existing RGT literature was conducted. Broadly, it seems that statistically significant choice interactions in RGT papers are often unsupported by visual inspection of the data. The simulations in the current project suggest many of these significant findings could be false positives. For example, there were significant dose by choice interactions in a repeated-measures ANOVA that examined the effects of disulfiram, a drug that affects both dopamine and norepinephrine, on RGT behavior (Di Ciano et al., 2018). However, all post hoc tests examining the dose effects on each individual RGT choice were non-significant. First, this highlights the importance of using the correct post hoc tests. This is notable given that 56% of brain and spinal cord injury researchers used incorrect post hoc tests in a review of 125 published articles (Burke et al., 2013). A second issue is that researchers may seek alternative analytic techniques to support statistically significant interactions when post hoc tests are non-significant. In the disulfiram paper, the authors

subdivided rats into “optimizers” and “sub-optimizers” based on their choice profiles. They found that 25 and 50 mg/kg of disulfiram increased advantageous choice (P1 and P2) for sub-optimizers only. In this case, visual inspection of the figures does corroborate the statistics, although the effect sizes were relatively small. In the papers described below, alternative data analytic techniques were used to find statistically significant results that were not corroborated by visual inspection of the data.

In another recent paper that assessed cue reactivity as a predictor of RGT choice, rats were divided into sign trackers (interacted more with conditioned stimuli associated with reinforcement; i.e., pressed levers that were extended prior to reinforcer delivery in operant chamber) or goal trackers (interacted more directly with reinforcer delivery; i.e., nosepoked in food hopper of operant chamber where reinforcers were delivered) (Swintosky, Brennan, Koziel, Paulus, & Morrison, 2021). A repeated-measures ANOVA with choice as a within-subject factor (roughly equivalent to the full LMER model in the current study) was used to determine that sign and goal trackers only differed on choice of one option on the RGT. However, correlations were also used to predict RGT performance using a metric of cue reactivity. The authors concluded that cue reactivity was predictive of RGT performance. The actual *r*-values for the correlations ranged between 0.2 and 0.3, and the plots of the data would likely be interpreted as “no correlation” if significant *p*-values did not accompany them. From these data, the authors concluded that sign-tracking may be a useful method for predicting vulnerability to pathological gambling in clinical populations. In this paper, an amphetamine challenge was also conducted. The statistics (repeated measures ANOVA with choice as a within-subjects factor) were used to assert that amphetamine decreased optimal choice and risky choice. However, most doses of amphetamine had unclear effects on choice, as visually demonstrated by small dose-level effects

with overlapping error bars. Although the small magnitude of effects was addressed in the discussion, the abstract states “amphetamine increased choices of a low-risk/low-reward option at the expense of optimal and high-risk choices” (Swintosky et al., 2021). To avoid overstating RGT findings, a more nuanced interpretation would be beneficial.

Similarly, in a paper assessing the effects of amphetamine on the mouse version of the RGT, a repeated-measures ANOVA (with choice as a within-subjects factor) showed that a high dose increased P1 and decreased P2 and P3 choice. However, the error bars across the high dose and saline were overlapping, and the drug reduced the number of trials by over 50% (van Enkhuizen, Geyer, & Young, 2013). The authors claimed these findings suggested that the RGT had translational validity for mouse models of drug-induced mania. Thus, some published findings on decision-making using the RGT, particularly with pharmacological manipulations, may be overstated. For RGT researchers that use linear models for data analysis, there are several strategies that should be used to prevent false positives. First, choice interactions must always be further inspected with post-hoc testing. If post hoc tests are non-significant, the use of additional techniques to explain the interaction (e.g., correlations) should not be performed except as exploratory analysis needing further study. Second, all results should be corroborated through visual inspection of the data. Lastly, more emphasis should be placed on effect sizes, rather than *p*-values. Several published papers found statistically significant effects of drugs on the RGT, but needed a more thorough discussion of the small effect size (Di Ciano et al., 2018; Silveira, Murch, Clark, & Winstanley, 2016; van Enkhuizen et al., 2013).

One strategy that has been used in published literature to account for interdependencies among options and low effect sizes is the use of a score variable as a single outcome in a repeated-measures ANOVA or intercept-only LMER (e.g., Daniel et al., 2017 Di Ciano, 2015).

The score variable is calculated as the difference score between "safe" choices (P1+P2) minus risky choices (P3+P4). The drawback of this approach is that it lacks power to differentiate between shifts in optimal and suboptimal choice. For example, after TBI, there is a decrease in optimal choice (P2) and an increase in suboptimal choice (P1) (Shaver et al., 2019). If P1 and P2 were collapsed together into a score variable, there might be no detectable effect of TBI. Recent work provided additional evidence that a frontal TBI effect could not be fully captured simply by dissociating between safe and risky choices. Rather, TBI seemed to reduce sensitivity to reinforcement on the RGT rather than increasing preference for risky choices (Vonder Haar, 2022a; in press). Therefore, collapsing the outcomes into safe versus risky choices may not be a useful strategy for analyzing the effects of TBI (and potentially other CNS manipulations). A more powerful strategy to account for all four choices simultaneously may be to treat the choice variable as a categorical outcome.

Limitations

In the current study, we found that a commonly-used intercept-only LMER model (and by extension, repeated measures ANOVA) was a poor strategy for RGT data analysis. Although we have proposed that a binomial or multinomial logistic regression may be superior, this hypothesis has not yet been empirically tested. Future studies should identify the superior method through Monte Carlo simulation and provide reproducible code/instructions for data analysis. The other major limitation is generalizability. First, the findings were task-specific, and second, statistical literacy and resistance to change may hinder methodological changes.

Generalizability

One limitation of the project is that the findings are only directly applicable to RGT users. However, these results still provide evidence that the broader practice in behavior analysis

of coercing discrete trials into a continuous variable is problematic. There are several common behavior analytic choice paradigms with categorical interdependent outcomes, such as the delay discounting task, where rodents choose between two levers, one of which provides a small, immediate reinforcer, and the other provides a larger but delayed reinforcer (Mazur, 1987). Other common tasks include discrimination, effort discounting, and reversal learning. In theory, the interdependencies on a 2-choice or 3-choice task should be exacerbated compared to a 4-choice task such as the RGT. Nonetheless, it is still standard practice to analyze choice at the aggregate level using linear models. This practice likely developed because non-linear regression was once too computationally intensive to perform with repeated-measures outcomes (Meddings et al., 1989).

Importantly, 2-choice outcomes can still be transformed and analyzed using linear models without violating the independence assumption if only one option is considered. However, these techniques may be less powered to detect effects compared to categorical analyses, which better fit the raw structure of the data. There are now accessible methods and software for relatively simple and efficient categorical analysis of repeated measures data (e.g., *glmer* in R; for example syntax, see the supplement for Young, 2018). Monte Carlo simulations have demonstrated that mixed-effects logistic regression outperformed linear regression for binary data, even when an arcsine-squareroot transformation was applied for the linear model (Dixon, 2008). Mixed-effects logistic regression has also been proposed as the best method for analyzing delay discounting data (Young, 2018). Thus, although this project does not directly translate to other behavior analytic paradigms, it contributes to a body of literature showing that choice data with discrete trials should be analyzed in raw form using categorical methods.

Choice paradigms with continuous, interdependent outcomes present a more complex problem. For example, time spent swimming, floating, versus escaping on the FST would be more difficult to analyze using a multinomial or binomial logistic regression. Technically, performance could be collapsed into a single value per subject, but this approach may reduce power and fail to capture the more continuous nature of the outcome variables. Outcomes could also be collapsed into a proportion of total time and analyzed using a weighted glmer model in the same fashion as the binomial mixed-effects regression in the current study. Thus, the findings here may generalize to a variety of other behavioral neuroscience tasks, and other analyses outside of ANOVA and linear regression should be explored for these tasks.

Another potential analytic technique for the FST and similar measures with interdependent outcomes stemming from continuous data is beta regression, which deals with proportion variables that are quasi-continuous because they are bound between 0 and 1 (Douma & Weedon, 2019). This technique may be more powerful than a binomial logistic regression because it is an extension of logit models specifically for responses continuous on the 0-1 interval, and thus, may be a better fit to analyzing variables like proportion of time spent swimming. However, beta regression use is sparse in preclinical literature; it has been used for some biological analyses (e.g., microbiome composition; Chai, Jiang, Lin, & Liu, 2018), but does not seem to be used for behavior analysis. Preclinical research would likely benefit from considering these alternative techniques rather than using linear models to analyze variables that do not have a truly linear relationship.

Another obstacle to generalizability is missing data. In the control set, missing data was minimal and primarily fell under the category of missing completely at random (e.g., missing due to an acute technical issue with an operant chamber). Data simulations were conducted

without any missingness. Thus, these findings might not generalize as well to data with high missingness or data that is systematically missing, particularly given that mixed-effects models are designed to handle data missing completely at random. Nonetheless, these RGT findings may be relevant for other choice paradigms used in behavioral neuroscience research to study psychiatric deficits. Identifying the best analytic technique for these various tasks remains an open question in the field. After identifying the best practices, the next major concern is implementing those practices.

Implementation

In addition to limitations of generalizability, the implementation of new analytic techniques is a major barrier to generating scientific impact. There are published papers from as early as the 1980s encouraging behavior analysts to use non-linear rather than linear models when analyzing dose-response curves (Meddings et al., 1989). Despite the growing body of literature demonstrating the advantages of non-linear models (and generalized linear models using nonlinear link functions) for preclinical choice paradigms (e.g., Dixon, 2008; Young, 2018), linear models on aggregate data remain the prevalent approach. This resistance to change is a major barrier to scientific advancement and is particularly pronounced for statistical methods. Reasons for resistance to statistical innovation include lack of awareness of recent developments, usability of statistical software, inadequate education, and lack of mandates for statistical rigor in publications (Sharpe, 2013). In the context of operant data analysis, linear regression is much more familiar to the average RGT user than a multinomial logistic regression, for example. Although multinomial logistic regression is theoretically the best approach for analyzing RGT data, it requires advanced statistical knowledge and may necessitate a Bayesian approach to incorporate repeated measures with mixed effects. There are published protocols for

Bayesian analysis in R, but many behavior analysts are unfamiliar with these methods and particularly resistant to the specification of priors (Young, 2019). Thus, it is important for future projects to determine if a simpler model, such as a mixed-effects binomial logistic regression is suitable for RGT data analysis.

Unfortunately, the sparse use of binomial logistic regression (or Poisson regression at the count level) for 2-choice tasks (e.g., delay discounting) suggests that implementation of new statistical techniques will be difficult. Advancements in statistics will greatly benefit the field of behavior analysis and help establish it as a valuable modern science. The survival of behavior analysis is particularly important for behavioral neuroscience due to the robust nature of behavior analytic methods for chronic measurement of psychiatric symptoms. Particularly, in the field of TBI, spatial learning measures, such as the Morris Water Maze dominate the field. These assays are less suited to repeated-measures testing and only capture a small subset of psychiatric symptoms caused by brain injury. By contrast, operant methods are more powerful for extended measurement of various psychiatric symptoms, including risky decision-making, motor/choice impulsivity, attention, and behavioral flexibility. For example, when rats were tested on various behavioral paradigms at 10-12 months after a frontal TBI, deficits were detected on differential reinforcement of low rate behavior (operant measure of impulsivity) but not on the rotarod task (non-operant sensorimotor task) (Lindner et al., 1998). Behavioral neuroscience benefits from the use of operant methods, and improvements in statistical methods may help narrow the translational gap between preclinical and clinical research.

Implications

In the current study, we identified substantial weaknesses in a common analytic approach to RGT data. The false positive rate was over 75% when analyzing choice of all four outcomes

as the dependent variable, and this was consistent across sample sizes. Some researchers have combatted this by analyzing a single score variable as the ratio of safe choices (P1 and P2) to risky choices (P3 and P4). Both approaches have major drawbacks that may hinder translation of RGT findings. The high false positive rate of the full LMER analysis suggests that at the preclinical level, there may be statistically significant findings that will inevitably fail to become clinically meaningful when translated. The score variable approach lacks power to detect subtle behavioral effects, such as the dissociation between suboptimal and optimal choice; however, this approach does translate more directly to the Iowa Gambling Task (IGT), which is used to measure similar constructs in humans.

There are a few obstacles to direct translation between the RGT and IGT. Some of these barriers are broadly applicable to a variety of translational tasks (e.g., inherent differences across species), and others are more task specific. One task-specific challenge is that the RGT and IGT can capture slightly different constructs. As discussed previously, the RGT can dissociate between optimal, suboptimal, and risky decisions, whereas the IGT only dissociates between optimal and risky decisions. This may be particularly difficult to reconcile because patients with large frontal brain lesions increased specifically in risky decisions on the IGT (Manes et al., 2002), whereas rats with a prefrontal brain injury had more robust increases in suboptimal but non-risky choice (Vonder Haar, 2022a; in press). It is difficult to discern whether IGT findings truly reflect a shift in risk preference as opposed to broader changes in the ability to discriminate between outcomes. The IGT may benefit from adding a suboptimal but non-risky option to better detect changes in discrimination. Furthermore, the RGT is also able to capture additional psychiatric deficits, such as motor impulsivity and psychomotor deficits, which cannot be detected by the IGT.

However, there are similarities across RGT and IGT research that suggest that preclinical findings can be useful. In particular, we see substantial variability across intact subjects on the RGT. The *k*-means clustering approach demonstrated that multiple phenotypes of non-optimal decision-makers exist, even among intact rats. This phenomenon is recapitulated in the clinical literature; there is considerable variability across individuals, and many healthy participants perform at an "impaired" level (Bull, Tippett, & Addis, 2015). The clustering approach in the current study may be useful for IGT researchers interesting in exploring individual subject variability. Another preclinical challenge with RGT research is determining whether shifts in behavior are driven by changes in risk preference or if results reflect a reduction in sensitivity to contingencies of reinforcement and punishment. We found that TBI rats have reduced sensitivity to reinforcement and may be less able to discriminate between outcomes (Vonder Haar, 2022a; in press). This question also applies to IGT research, and alternative decision-making tasks have been used to determine whether deficits on the IGT are reflective of reduced sensitivity to reinforcement and punishment. In a study of healthy participants, individuals that performed poorly on the IGT did have a reduced sensitivity to magnitude of reinforcement and punishment (Bull et al., 2015). Thus, there are many unique questions that can be answered using both the RGT and IGT and compared across the two. The similarities between the RGT and IGT also provide an excellent opportunity to glean insights into individual subject variability and identify risk and resilience factors and potential therapeutics. However, inappropriate statistical techniques may continue to hinder translation, and future work may be more likely to translate if more accurate data analytic techniques are used at the preclinical level.

Future Directions

The current approaches for RGT data analyses are flawed, and it will be important to identify superior techniques via data simulation in future studies. Specifically, binomial and multinomial logistic regression should be compared to determine the power of each test for RGT analysis. It may also be prudent to test the accuracy of LMER with additional sessions and with random slopes. Given the constraints of convincing the field to adopt new techniques, the "best" method may not be the method with the highest accuracy; rather, it is important to balance both accuracy and feasibility. For example, a multinomial logistic regression using the *brms* package is time-consuming and requires skills that are unfamiliar to most behavioral researchers. Fitting the Bayesian multinomial regression for a single dataset (n = 20 per group) in the current study took approximately 20 minutes. Notably, this analysis was performed on proportion level data, which reduces computing time. If a researcher wanted to assess learning effects within a session across individual trials, the computing time would increase even further. If a researcher has one specific model to run, it might be reasonable to perform the Bayesian multinomial logistic regression; however, the computation time may grow prohibitively lengthy for multiple model comparisons. Further, any errors in the process may take considerable time to isolate. Because of these obstacles, a more familiar and less computationally intensive analysis (i.e., binomial logistic regression) may be more desirable. If results across both binomial and multinomial analyses are reasonably similar, it would be advantageous to promote the technique that is easier to implement.

In addition to eventual publication of these analyses in a peer-reviewed journal, it is important to disseminate results through other forums. There have recently been calls for more transparency in science, with the NIH stating that "data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential

and proprietary data" (Health, 2020). Given that published methods should be reproducible, sharing code and data (when ethical) are essential for statistical advancement. The code for this project will be shared at https://github.com/mfrankz at the time of publication. When the best method for RGT data analysis is identified in future projects, that code will also be shared freely to promote the adoption of accurate statistical practices. Forums such as GitHub allow for easy sharing of code and will continue to play a major role in scientific advancement.

In addition to these actions that can be taken by individual researchers, there are also more systemic changes that would facilitate better statistical practices. First, journals should implement and expand requirements for data sharing and code availability (Walters, 2020). Second, the curriculum for statistics classes in psychology departments should be evaluated to determine if students are gaining an accurate understanding of statistics and their limitations. The focus on null-hypothesis testing and surface-level skills (e.g., memorizing formulas, following instructions in SPSS) without an emphasis on critical thinking may promote the implementation of poor statistical practices and discourage scientific advancement. Although these types of changes can be slow to implement, it is important to note that behavioral neuroscientists are frequently eager to adopt the most advanced techniques in the field. Several techniques recently considered novel (e.g., optogenetics, single-cell sequencing, advanced microscopy) have been quickly and enthusiastically adopted. This same desire for innovation should be applied to statistical techniques as well. It may be challenging to adopt more accurate statistical practices, but it will likely improve translation and ultimately benefit the field.

Concluding Remarks

Poor translation between preclinical and clinical research is one of the most consequential problems in behavioral neuroscience. For RGT researchers specifically, the use of inappropriate

statistical techniques has likely resulted in both false positives and reduced power. Although the current study only provided evidence for inaccuracies in RGT analysis, these problems are most likely pervasive across many preclinical choice paradigms. Statistical practices must evolve to improve the accuracy of preclinical data analysis and narrow the gap in translation. Simulation projects are crucial for identifying the best statistical practices. Dissemination of these practices presents a more complex issue and will require publication, open code sharing, and critical reflection on how statistical techniques should be taught to students and early-career researchers.

References

Antunes, M., & Biala, G. (2012). The novel object recognition memory: neurobiology, test procedure, and its modifications. *Cognitive processing, 13*(2), 93-110. doi:10.1007/s10339-011-0430-z

Bahceci, D., Anderson, L. L., Occelli Hanbury Brown, C. V., Zhou, C., & Arnold, J. C. (2020). Adolescent behavioral abnormalities in a Scn1a(+/-) mouse model of Dravet syndrome. *Epilepsy Behav, 103*(Pt A), 106842. doi:10.1016/j.yebeh.2019.106842

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*(3). doi:10.1016/j.jml.2012.11.001

Barrus, M. M., Hosking, J. G., Zeeb, F. D., Tremblay, M., & Winstanley, C. A. (2015). Disadvantageous decision-making on a rodent gambling task is associated with increased motor impulsivity in a population of male rats. *J Psychiatry Neurosci, 40*(2), 108-117. doi:10.1503/jpn.140045

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *arXiv, 1506*.

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

Baum, W. M. (1989). Quantitative prediction and molar description of the environment. *The Behavior analyst, 12*(2), 167-176. doi:10.1007/BF03392493

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15. doi:10.1016/0010-0277(94)90018-3

Begg, C. B., & Gray, R. (1984). Calculation of Polychotomous Logistic Regression Parameters Using Individualized Regressions. *Biometrika, 71*(1), 11-18. doi:10.2307/2336391

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol Methods, 17*(3), 399-417. doi:10.1037/a0028087

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi:https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Bragge, P., Synnot, A., Maas, A. I., Menon, D. K., Cooper, D. J., Rosenfeld, J. V., & Gruen, R. L. (2016). A State-of-the-Science Overview of Randomized Controlled Trials Evaluating Acute Management of Moderate-to-Severe Traumatic Brain Injury. *J Neurotrauma, 33*(16), 1461-1478. doi:10.1089/neu.2015.4233

Broadbent, N. J., Gaskin, S., Squire, L. R., & Clark, R. E. (2009). Object recognition memory and the rodent hippocampus. *Learning & memory (Cold Spring Harbor, N.Y.), 17*(1), 5-11. doi:10.1101/lm.1650110

Bruijnzeel, A. W., Knight, P., Panunzio, S., Xue, S., Bruner, M. M., Wall, S. C., . . . Setlow, B. (2019). Effects in rats of adolescent exposure to cannabis smoke or THC on emotional behavior and cognitive function in adulthood. *Psychopharmacology (Berl), 236*(9), 2773-2784. doi:10.1007/s00213-019-05255-7

Bull, P. N., Tippett, L. J., & Addis, D. R. (2015). Decision making in healthy participants on the Iowa Gambling Task: new insights from an operant approach. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00391

Burke, D. A., Whittemore, S. R., & Magnuson, D. S. K. (2013). Consequences of common data analysis inaccuracies in CNS trauma injury basic research. *Journal of neurotrauma, 30*(10), 797-805. doi:10.1089/neu.2012.2704

Bürkner, P. C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395-411. doi:10.32614/RJ-2018-017

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Stat Med, 25*(24), 4279-4292. doi:10.1002/sim.2673

Chai, H., Jiang, H., Lin, L., & Liu, L. (2018). A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput Biol, 14*(7), e1006329. doi:10.1371/journal.pcbi.1006329

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. . (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.): Routledge.

Cole, E., Ziadé, J., Simundic, A., & Mumby, D. G. (2020). Effects of perirhinal cortex and hippocampal lesions on rats' performance on two object-recognition tasks. *Behav Brain Res, 381*, 112450. doi:10.1016/j.bbr.2019.112450

Daniel, M. L., Cocker, P. J., Lacoste, J., Mar, A. C., Houeto, J. L., Belin-Rauscent, A., & Belin, D. (2017). The anterior insula bidirectionally modulates cost-benefit decision-making on a rodent gambling task. *Eur J Neurosci, 46*(10), 2620-2628. doi:10.1111/ejn.13689

Davison, M., & Baum, W. M. (2002). Choice in a variable environment: effects of blackout duration and extinction between components. *Journal of the experimental analysis of behavior, 77*(1), 65-89. doi:10.1901/jeab.2002.77-65

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876-879. doi:10.1038/nature04766

Di Ciano, P., Manvich, D. F., Pushparaj, A., Gappasov, A., Hess, E. J., Weinshenker, D., & Le Foll, B. (2018). Effects of disulfiram on choice behavior in a rodent gambling task: association with catecholamine levels. *Psychopharmacology (Berl), 235*(1), 23-35. doi:10.1007/s00213-017-4744-0

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447-456. doi:https://doi.org/10.1016/j.jml.2007.11.004

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution, 10*(9), 1412-1430. doi:https://doi.org/10.1111/2041-210X.13234

Dutech, A., Coutureau, E., & Marchand, A. R. (2011). A reinforcement learning approach to instrumental contingency degradation in rats. *J Physiol Paris, 105*(1-3), 36-44. doi:10.1016/j.jphysparis.2011.07.017

Dwivedi, S. B., L. (2019). *A Systematic Review on K-Means Clustering Techniques*.

Ennaceur, A., & Delacour, J. (1988). A new one-trial test for neurobiological studies of memory in rats. 1: Behavioral data. *Behav Brain Res, 31*(1), 47-59. doi:10.1016/0166-4328(88)90157-x

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *Am Psychol, 63*(7), 591-601. doi:10.1037/0003-066x.63.7.591

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. In C. a. Hall (Ed.), (Second Edition ed.). doi:https://doi.org/10.1201/9781315382722

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Garner, J. P. (2014). The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *Ilar j, 55*(3), 438-456. doi:10.1093/ilar/ilu047

Health, N. I. o. (2020). NIH Data Sharing Policy and Implementation Guidance
. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Heisig, J., & Schaeffer, M. (2018). *Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction*.

Henry, L., Wickham, H. (2020). purrr: Functional Programming Tools. R package version 0.3.4. Retrieved from https://CRAN.R-project.org/package=purrr

Herrnstein, R. J. (1970). On the law of effect. *Journal of the experimental analysis of behavior, 13*(2), 243-266. doi:10.1901/jeab.1970.13-243

Hoekstra, R., Kiers, H. A., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Front Psychol, 3*, 137. doi:10.3389/fpsyg.2012.00137

Hornoiu, I., Gigg, J., & Talmi, D. (2020). Quantifying how much attention rodents allocate to motivationally-salient objects with a novel object preference test. *Behav Brain Res, 380*, 112389. doi:10.1016/j.bbr.2019.112389

Knief, U., & Forstmeier, W. (2020). Violating the normality assumption may be the lesser of two evils. *bioRxiv*, 498931. doi:10.1101/498931

Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *WIREs Computational Statistics, 6*(6), 386-392. doi:https://doi.org/10.1002/wics.1314

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974.

Lenth, R. V. (2021). Estimated Marginal Means, aka Least-Squares Means. Retrieved from https://CRAN.R-project.org/package=emmeans

Lindner, M. D., Plone, M. A., Cain, C. K., Frydel, B., Francis, J. M., Emerich, D. F., & Sutton, R. L. (1998). Dissociable long-term cognitive deficits after frontal versus sensorimotor cortical contusions. *Journal of neurotrauma, 15*(3), 199-216. doi:10.1089/neu.1998.15.199

Luce. (1959). *Individual choice behavior: A theoretical analysis*. Mineola, NY, US: Dover Publications.

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., & Robbins, T. (2002). Decision-making processes following damage to the prefrontal cortex. *Brain, 125*(Pt 3), 624-639. doi:10.1093/brain/awf049

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In *The effect of delay and of intervening events on reinforcement value.* (pp. 55-73). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

McCullagh, P. N., J.A. . (1989). *Generalized Linear Models*: Chapman and Hall.

Meaney, C., & Moineddin, R. (2014). A Monte Carlo simulation study comparing linear regression, beta regression, variable-dispersion beta regression and fractional logit regression at recovering average difference measures in a two sample design. *BMC Med Res Methodol, 14*, 14. doi:10.1186/1471-2288-14-14

Meddings, J. B., Scott, R. B., & Fick, G. H. (1989). Analysis and comparison of sigmoidal curves: application to dose-response data. *Am J Physiol, 257*(6 Pt 1), G982-989. doi:10.1152/ajpgi.1989.257.6.G982

Mezadri, T. J., Batista, G. M., Portes, A. C., Marino-Neto, J., & Lino-de-Oliveira, C. (2011). Repeated rat-forced swim test: reducing the number of animals to evaluate gradual effects of antidepressants. *J Neurosci Methods, 195*(2), 200-205. doi:10.1016/j.jneumeth.2010.12.015

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi:10.1037/0033-2909.105.1.156

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol, 7*, 34. doi:10.1186/1471-2288-7-34

Moreton, E., Baron, P., Tiplady, S., McCall, S., Clifford, B., Langley-Evans, S. C., . . . Voigt, J. P. (2019). Impact of early exposure to a cafeteria diet on prefrontal cortex monoamines and novel object recognition in adolescent rats. *Behav Brain Res, 363*, 191-198. doi:10.1016/j.bbr.2019.02.003

Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*(2), 239-260. doi:https://doi.org/10.1016/0023-9690(81)90020-5

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074-2102. doi:https://doi.org/10.1002/sim.8086

Munyon, C., Eakin, K. C., Sweet, J. A., & Miller, J. P. (2014). Decreased bursting and novel object-specific cell firing in the hippocampus after mild traumatic brain injury. *Brain Res, 1582*, 220-226. doi:10.1016/j.brainres.2014.07.036

Namiki, N., Oyo, K., & Takahashi, T. (2015). *How Do Humans Handle the Dilemma of Exploration and Exploitation in Sequential Decision Making?*

Porsolt, R. D., Anton, G., Blavet, N., & Jalfre, M. (1978). Behavioural despair in rats: a new model sensitive to antidepressant treatments. *Eur J Pharmacol, 47*(4), 379-391. doi:10.1016/0014-2999(78)90118-8

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rogers, J., Churilov, L., Hannan, A. J., & Renoir, T. (2017). Search strategy selection in the Morris water maze indicates allocentric map formation during learning that underpins spatial memory formation. *Neurobiol Learn Mem, 139*, 37-49. doi:10.1016/j.nlm.2016.12.007

Seyhan, A. A. (2019). Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications, 4*(1), 18. doi:10.1186/s41231-019-0050-7

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *18*, 572-582. doi:10.1037/a0034177

Shaver, T. K., Ozga, J. E., Zhu, B., Anderson, K. G., Martens, K. M., & Vonder Haar, C. (2019). Long-term deficits in risky decision-making after traumatic brain injury on a rat analog of the Iowa gambling task. *Brain Res, 1704*, 103-113. doi:10.1016/j.brainres.2018.10.004

Shimp, C. P. (2020). Molecular (moment-to-moment) and molar (aggregate) analyses of behavior. *Journal of the experimental analysis of behavior, 114*(3), 394-429. doi:10.1002/jeab.626

Silveira, M. M., Murch, W. S., Clark, L., & Winstanley, C. A. (2016). Chronic atomoxetine treatment during adolescence does not influence decision-making on a rodent gambling task, but does modulate amphetamine's effect on impulsive action in adulthood. *Behav Pharmacol, 27*(4), 350-363. doi:10.1097/fbp.0000000000000203

Sivakumaran, M. H., Mackenzie, A. K., Callan, I. R., Ainge, J. A., & O'Connor, A. R. (2018). The Discrimination Ratio derived from Novel Object Recognition tasks as a Measure of Recognition Memory Sensitivity, not Bias. *Scientific Reports, 8*(1), 11579. doi:10.1038/s41598-018-30030-7

Slattery, D. A., Desrayaud, S., & Cryan, J. F. (2005). GABAB receptor antagonist-mediated antidepressant-like behavior is serotonin-dependent. *J Pharmacol Exp Ther, 312*(1), 290-296. doi:10.1124/jpet.104.073536

Stopper, C. M., & Floresco, S. B. (2011). Contributions of the nucleus accumbens and its subregions to different aspects of risk-based decision making. *Cogn Affect Behav Neurosci, 11*(1), 97-112. doi:10.3758/s13415-010-0015-9

Sutton, R. S., & Barto, A.G. (1998). *Reinforcement Learning*: The MIT Press.

Swintosky, M., Brennan, J. T., Koziel, C., Paulus, J. P., & Morrison, S. E. (2021). Sign tracking predicts suboptimal behavior in a rodent gambling task. *Psychopharmacology (Berl), 238*(9), 2645-2660. doi:10.1007/s00213-021-05887-8

Vaishnavi, S., Rao, V., & Fann, J. R. (2009). Neuropsychiatric problems after traumatic brain injury: unraveling the silent epidemic. *Psychosomatics, 50*(3), 198-205. doi:10.1176/appi.psy.50.3.198

van Enkhuizen, J., Geyer, M. A., & Young, J. W. (2013). Differential effects of dopamine transporter inhibitors in the rodent Iowa gambling task: relevance to mania. *Psychopharmacology (Berl), 225*(3), 661-674. doi:10.1007/s00213-012-2854-2

Vonder Haar, C., Frankot, M., Reck, A., Milleson, V., & Martens, K. (2022a). Large-N rat data enables phenotyping of risky decision-making: A retrospective analysis of brain injury on the Rodent Gambling Task. *Frontiers in Behavioral Neuroscience*.

Vonder Haar, C., Martens, K. M., Frankot, M. A. . (2022b). *Combined dataset of Rodent Gambling Task in rats after brain injury*.

Vorhees, C. V., & Williams, M. T. (2006). Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nat Protoc, 1*(2), 848-858. doi:10.1038/nprot.2006.116

Walters, W. P. (2020). Code Sharing in the Open Science Era. *Journal of Chemical Information and Modeling, 60*(10), 4417-4420. doi:10.1021/acs.jcim.0c01000

Young, M. E. (2018). Discounting: A practical guide to multilevel analysis of choice data. *Journal of the experimental analysis of behavior, 109*(2), 293-312. doi:10.1002/jeab.316

Young, M. E. (2019). Bayesian data analysis as a tool for behavior analysts. *Journal of the experimental analysis of behavior, 111*(2), 225-238. doi:10.1002/jeab.512

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*(2), 160-177. doi:10.1016/j.lmot.2008.10.004

Young, M. E., Cole, J. J., & Sutherland, S. C. (2012). Rich stimulus sampling for between-subjects designs improves model selection. *Behav Res Methods, 44*(1), 176-188. doi:10.3758/s13428-011-0133-5

Young, M. E., & Hoane, M. R. (2021). Mixed effects modeling of Morris water maze data revisited: Bayesian censored regression. *Learning & behavior, 49*(3), 307-320. doi:10.3758/s13420-020-00457-y

Zeeb, F. D., & Winstanley, C. A. (2013). Functional disconnection of the orbitofrontal cortex and basolateral amygdala impairs acquisition of a rat gambling task and disrupts animals' ability to alter decision-making behavior after reinforcer devaluation. *J Neurosci, 33*(15), 6434-6443. doi:10.1523/jneurosci.3971-12.2013