Investigating the effects of common analytical techniques on reaction time data

by

Angela Crumer

B.S., Southeast Missouri State University, 2008
M.S., Kansas State University, 2011

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

The heavy right skew of reaction time data creates challenges for analyses. Common analytical techniques may require a set of assumptions that are not found in this type of data. Some of the effects are known while others are not. The current study uses Monte Carlo simulation to assess the effects of ignoring distributional assumptions, aggregation, transformation, and truncation on reaction time data. The effects of these current practices were compared to fitting a generalized linear model. Each analysis was simulated to obtain false alarm and hit rates. From these values, the discriminability and criterion values from signal detection theory were calculated. Parameter estimates were also obtained and compared to the theoretical values from the simulation to produce estimates of parameter bias and accuracy. While fitting a generalized linear model had the highest discriminability and unbiased criterion, it was not very different from ignoring distributional assumptions and aggregating the data. Transforming the data using a log transformation resulted in biased and inaccurate parameter estimates and had the lowest discriminability. Truncating the data inflated the error and resulted in poor signal detection and poor parameter estimation.

Investigating the effects of common analytical techniques on reaction time data

by

Angela Crumer

B.S., Southeast Missouri State University, 2008
M.S., Kansas State University, 2011

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Michael Young

Abstract

The heavy right skew of reaction time data creates challenges for analyses. Common analytical techniques may require a set of assumptions that are not found in this type of data. Some of the effects are known while others are not. The current study uses Monte Carlo simulation to assess the effects of ignoring distributional assumptions, aggregation, transformation, and truncation on reaction time data. The effects of these current practices were compared to fitting a generalized linear model. Each analysis was simulated to obtain false alarm and hit rates. From these values, the discriminability and criterion values from signal detection theory were calculated. Parameter estimates were also obtained and compared to the theoretical values from the simulation to produce estimates of parameter bias and accuracy. While fitting a generalized linear model had the highest discriminability and unbiased criterion, it was not very different from ignoring distributional assumptions and aggregating the data. Transforming the data using a log transformation resulted in biased and inaccurate parameter estimates and had the lowest discriminability. Truncating the data inflated the error and resulted in poor signal detection and poor parameter estimation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1- introduction

Investigating the Effects of Common Analytical Techniques on Reaction Time Data

A growing niche in cognitive psychology is model fitting, simply called modeling. Modeling refers to fitting the appropriate model to a set of data. This requires estimating parameter values from a sample and fitting them to a theoretical population distribution function. Many techniques exist to compare model fits to one another and determine the best fit model.

One area of research that is particularly interested in model fitting is reaction time (RT) research. Reaction time data refers to data in which the dependent variable (DV) is the time until a response is made. A common example is visual search, where subjects are shown an image and asked to make simple decisions and respond (Palmer et al, 2011). In a search task, a subject is asked to respond when a particular target is present and withhold responding when it is absent. The DV, in this example, is how long it takes to make a response. These reaction times tend to be relatively short, around 800-1200 ms. Another example is decision making. A subject may be asked to select between two images (e.g., which is preferred or which is a member of a category) or determine if two images are similar or different. Here the DV is the time it takes to make a decision. These reaction times tend to be a little longer because a decision making process will be employed, often as long as 1500 ms. Another example in which RTs are recorded and analyzed is eye tracking. A subject is shown an image while their fixation location is tracked. The DV is the time it takes for the eye to move between positions, sometimes called latency. These reaction times tend to be extremely short, as short as 200 ms. Because RT data measures time, it has a lower bound of 0 but no upper bound, causing RT data to be heavily right skewed (Fig 1). This makes analyzing reaction time data more complex because many of the common

analytical techniques, such as analysis of variance (ANOVA) and linear regression, assume an underlying normal distribution.

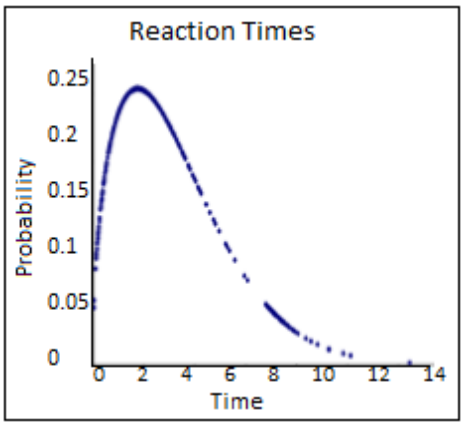

Figure 1. Reaction times. This figure shows simulated reaction time data.

Historically, researchers have approached the analysis of RT data in multiple ways, some more appropriate than others. Although modeling RT data with the various techniques is common, the varying consequences of using these common techniques are not well known. What exactly is affected when researchers do not pay attention to the underlying distribution of the data? What is lost when aggregating and could this technique be beneficial under some conditions? What are some of the costs of transforming or truncating data? The current project targets these questions regarding common analytical techniques for RT data in an effort to provide a more comprehensive understanding of the relative strengths and weaknesses of these techniques across situations.

To begin, I will discuss the assumptions, uses, and issues with analytical techniques commonly used for analyzing RT data. In particular, I will address ignoring distributional assumptions, aggregation, transformations, and truncation, followed by a comparison of relative strengths and weaknesses of these techniques. Next, I describe and discuss Monte Carlo simulations and signal detection theory, and how they are used to assess the aforementioned

techniques. I continue with a discussion about generalized linear modeling, the Gamma distribution and how it compares with the other common analytical techniques. Finally, I present preliminary findings and specific hypotheses regarding the discriminability, criterion, and parameter recovery of each of the 5 techniques for analyzing RT data.

Ignoring distributional assumptions

A common analytical technique involves ignoring the distributional assumptions and proceeding with techniques that assume normally distributed residuals. In these analyses, no attempt is made to correct for the skew in the data. ANOVA or linear regression is applied to the raw data, even though the normality assumption has been violated. Many times researchers claim this technique is robust to violations of normality because of the properties of the central limit theorem (CLT). The CLT states that the distribution of sample means will be approximately normal for large sample sizes. This is entirely different than claiming the distribution of a particular sample (and underlying population distribution) will be approximately normal for a large sample size. Therefore, making this claim does not satisfy the distributional assumptions; it ignores them. Ignoring distributional assumptions can be detrimental to the analysis because it inaccurately describes the variation in the data as being symmetrically distributed around a mean value. In right skewed data, variation around smaller values will be much smaller than variation around larger values because of the floor effect. This creates heterogeneity of variance, thus violating another assumption of standard general linear modeling techniques. Misattributing error leads to several misleading conclusions. It can cause the patterns of means to be distorted (Wolfe, Palmer & Horowitz, 2010), create misspecifications of interactions in a model, and produce an increase in error rates (Dixon, 2008; Erceg-Hurn & Mirosevich, 2008; Lorch & Myers, 1990). These errors, though separate and unique, go hand in hand. It may appear that condition means are shifting, when in fact they are simply becoming more variable as they move

away from the lower bound (i.e., RT increases). As the patterns of means becomes increasingly distorted, strange results occur within the analyses, such as unrealistic interactions (Dixon, 2008). This heterogeneity of variances will increase the likelihood of Type I error, indicating an interaction effect that is not present in the data because the changing variability in the means is not accounted for.

Aggregation

Another technique used to simplify the data for analysis is aggregation. Many researchers collect reaction time data over many trials, and it is common to average across trials to get a mean RT score for each subject with these means subsequently entered into a linear regression model or ANOVA. In the case of RT data, the mean is not the appropriate measure of central tendency because it is affected by the extreme observations in the right tail. Sometimes the median is used instead of the mean, because it is a better measure of central tendency for such data. But, the median varies more than other sample statistics (Ratcliff, 1993), and is a biased estimator (Miller, 1998), tending to overestimate the value of the true median. Therefore, it is also not a good measure of central tendency for model fitting.

Furthermore, when RTs are aggregated across trials, information from trial-to-trial variability is eliminated. In essence, trial becomes a fixed variable because all information about the variability in RTs across trials is lost (Lorch & Myers, 1990). Ensuring that behavior does not systematically change between trials would simplify analyses, but because trials occur over time, that is not the case. Subjects change attention over time (Posner, Snyder & Davidson, 1980), learn and adapt over time, or become fatigued over time. When data are aggregated across trial, none of these effects can be measured, let alone be accounted for. Furthermore, information about the sample size that produced the aggregate is lost. An analysis providing only the aggregate as an outcome variable will produce the same results whether the aggregate was based

on two RTs or two hundred. Aggregating data has been shown to cause an analysis to find more significant results than truly exist because it hides variability across measures (Lorch & Myers, 1990).

Transformations

It is also common for researchers to transform reaction times to increase the normality of the residuals. This is done in an attempt to make applying traditional methods, which assume normality, more appropriate. Commonly used transformations for reaction time data are logarithmic, inverse, and square root transforms (Bartlett, 1947). There are three known issues with transformations: some transformations may not reestablish normality (Baayen & Milin, 2010), they can make it difficult to interpret the results within the transformed space, and they can create complications when trying to convert back to the original scale (Bartlett, 1947; Dixon, 2008; Erceg-Hurn & Mirosevich, 2008). Applying transformations (i.e. log, inverse, etc.) also changes the original ratio scale, which can distort the interval differences. When time is the dependent variable, it is best to analyze raw RTs. If the RTs are transformed, the direct connection with mental processing is lost (Lo & Andrews, 2015).

Furthermore, with commonly used transformations, multiplicative effects are often masked when trying to back-transform the results. For example, a dependent variable may be the multiplicative combination of its predictor variables, but when log transformed that effect becomes additive due to the law of logarithms. This is a serious problem when assessing interactive effects. However, transformations like the popular logarithmic (Van Breukelen, 2005) help minimize the effects of outliers and reduce the unequal variability of the IV across the range of the DV (heteroscedasticity) (Bartlett, 1947). Although transforming dependent variables helps create a variable which can be analyzed using linear-scale methods, the question of usefulness remains if these methods hide the nature of the effect. Lo and Andrews (2015) reported

drastically different patterns in transformed data using the inverse transformation. In an experiment using lexical data, raw scores demonstrated an additive effect. These results would suggest independent, serial processing between predictors. In this experiment, the inverse transformation most effectively restored normality of the residuals. When the data were inverse-transformed, the results demonstrated under-additive patterns. These results suggest interactions between predictors, which is in direct opposition to the analysis of the raw data.

Truncation

Because researchers often notice that the long right tail is creating problems for a conventional analysis, they sometimes resort to truncating these "outliers" in order to increase statistical power. Because of the heavy right skew, distinguishing between an outlier and a valid but long RT can be challenging. An outlier is a response generated by some process other than the one of interest (e.g., experimenter error or inattention to task) and thus should be eliminated, but valid yet extreme values should be retained and properly accounted for in the statistical model. In RT data, most outliers will occur in the upper tail of the distribution, hidden among the other very long RTs (Ratcliff, 1993; Baayen & Milin, 2010). A common method of eliminating outliers in RT data is truncation. Researchers my omit observations that are unrealistically too short, such as RTs that are shorter than is physically possible. These RTs would not be included because they represent an anticipatory response instead of a reaction to the task. More commonly, researchers remove observations that are unusually long. Common truncation methods involve using an a priori range of acceptable RTs (Ulrich & Miller, 1994). The most common technique is eliminating data that are 2 to 3 standard deviations beyond the mean. However, because the mean is not a good measure of central tendency for skewed data, it is also not a good measure for deciding outlying responses. Because of the skew in the data, truncation will cause more observations in the upper tail to become eliminated than from the lower tail. In

most cases, lower truncation does not even occur. This creates biases when estimating parameters and fitting models to the data. Truncating observations in the tails may lead to the over exclusion of influential data values and possibly under exclusion of others (Baayen & Milin, 2010; Ratcliff, 1993). While outliers add noise to the data, the bias caused by truncation is more harmful (Ulrich & Miller, 1994). An alternative to truncation is censoring. In the case of reaction time data, censoring could result in the experimenter not recording responses made before a pre-established starting point to eliminate responses due to anticipatory mechanisms. It could also result in not recording responses made after a pre-established deadline to eliminate responses due to shifts in attention. Even though the full information on the censored responses is not available, it is still less harmful to the analysis than truncation (Dolan, van deer Maas & Molenaar, 2002).

Comparing the Relative Strengths and Weaknesses of Common Approaches

Each common analytical technique has its benefits and drawbacks. The most common benefit of these techniques is simplicity. Aggregating across trials or even subjects can drastically simplify an analysis. In some specific fields, such as organizational climate research, aggregation may be the preferred analytical technique. A researcher may wish to aggregate across individuals to get a measure of climate at the team, unit, or organization. However, group level RTs are not usually of interest and aggregation is most often chosen for simplicity rather than theory. When data are not normally distributed, truncating 'problematic' outliers or transforming data and applying normal models is simpler than determining the underlying distribution of the data and applying a generalized multilevel model. The simplicity of these mistakes may not be worth the problems they create. When analyses indicate non-existent differences between groups (i.e. false alarms) and provide biased parameter estimates, one

begins to question the cost of this added simplicity. However, the magnitude of these problems and the conditions under which they are more or less problematic is the issue addressed in this study. If the sample size is sufficiently large, does it matter if the analysis is not strictly correct? Do some of these approaches err on the side of being too conservative whereas others are too liberal? Are there specific situations where one or more of these techniques should be strictly avoided?

It is important for researchers to use appropriate techniques which lead to more accurate results, instead of preferring techniques for simplicity. If certain techniques are inflating Type 1 error (probability of rejecting H_o when H_o is true), it becomes harder to replicate those results. In the end, published results will be considered valid and will contribute to theory development but will not actually provide anything substantive if the Type 1 error is inflated. It is important that research maintains an appropriate level of Type 1 error, usually no more than 5%, so that subsequent theories and postulates are accurate and truthful. Additionally, researchers may settle on a significant result, believing it to be accurate, and may not continue the research to find any other actual results. On the other hand, using inappropriate analytical techniques may increase Type 2 error (probability of failing to reject H_o when H_o is false). Many researchers consider this error less problematic than a Type 1 error, although it depends on the nature of the research. In some areas, such as intervention or treatment development, missing significant results is worse than misidentifying non-significant results. Type 2 errors also hinder the progress of scientific development. Researchers discovering a non-significant result will likely move on to another line of study, instead of correcting the analytical techniques. They will try new variables, new conditions, and new subjects; all the while missing the actual significant result due to analytical errors. Even when Type 1 and Type 2 errors are controlled, using inappropriate analytical

techniques can result in mis-estimation of the parameters. Sometimes a more sophisticated model may not be the most appropriate. A mis-specified model can cause mis-estimation as well.When parameters are over-estimated, effects and differences may be thought to be more consequential than they are. Similarly, when parameters are under-estimated, important effects and differences may be missed. Either of these mis-estimation errors results in faulty conclusions and less replicable results. If results are not reproducible, no matter how rigorous the experiment or how established the originator may be, they cannot be supported.

Different techniques estimate parameters in an attempt to recover the actual parameter values (i.e. parameter recovery). The goal of the current research is to investigate the effects of different types of common analytical techniques used with reaction time data. I will compare Type 1 and Type 2 error rates, along with parameter recovery, to determine the appropriateness and replicability of commonly used techniques in the field today.

Monte Carlo Simulations

One way to test the effects that different analytical techniques have on RT data is to use Monte Carlo simulations (Metropolis & Ulam, 1949). Data are simulated with properties that mimic those of the datasets of interest, as shown in Figure 1. As part of the simulation, the researcher specifies the parameters in the model so they are known quantities. Sample statistics, such as mean and variability, are specified. Distributional properties and other statistics, such as effect size are also set. As part of the simulation, the researcher can alter the sample size, number of trials per subject, number of subjects per condition, number of conditions, etc. The type of variables (continuous or discrete) and any dependent relationships between variables or responses are also part of the simulation. Through this technique, the researcher simulates the exact conditions that he or she wishes to study. Different modeling techniques can be used with

the simulated data that have differing properties. In this manner, it is possible to identify the conditions under which some techniques are particularly problematic and conditions under which they are not.

Once an adequate sample has been simulated, different analytical techniques are applied to the simulated data. Because the researcher specified the different parameters in the simulation, a direct comparison can be made between the estimates from the analyses and the true values. Verifying the accuracy of an analytical technique in predicting the parameters in the data becomes possible. Techniques that estimate values close to those specified in the simulation are preferred to techniques that do not. With the use of Monte Carlo simulations, researchers are also able to consider a variety of different conditions, such as differing sample or effect sizes. Determining which techniques work best under certain conditions will help inform the field of appropriate ways to analyze data. It is also useful in determining the strengths and weaknesses of different models and techniques. Some techniques may work well for non-normal data, but break down for small samples (Oberfeld & Franke, 2013). Other techniques could produce consistent results regardless of sample size, but may only be effective for large effect sizes. Because all of these conditions are able to be specified through simulations, the researcher can get a more global understanding about the nature of different analytical techniques.

Most importantly, with the use of simulation studies, one can more readily assess the specific effects of an analytical technique on the results. With simulation studies, it has been discovered that some techniques, such as aggregation, affect the results by reducing variability and increasing Type 1 error (Lorch & Myers, 1990). Other techniques, such as truncation, affect the results by adding bias in the parameter estimates (Ulrich & Miller, 1994).

Signal Detection Theory

Signal detection theory (SDT) is a unique way to measure the effectiveness of different modeling techniques. The constructs from SDT can be quantified for each technique and ultimately be used to differentiate the efficacy of the different techniques. This theory describes decision making as a process of discrimination between noise and signal. Noise is irrelevant information and signal is the actual event of interest. In simulation studies, SDT is applied to identifying effects. Identifying a true effect means correctly rejecting the null hypothesis, much like correctly identifying the signal. Signal detection theory describes the relationship between these two factors.

Being able to differentiate between noise and signal depends on how similar the noise is to the signal. If the noise is very similar to the signal, it will be more difficult to tell them apart and more mistakes will be made. If the irrelevant information is very different from the signal, differentiation will be easier and fewer mistakes will be made. For example, consider a participant sampling different kinds of soda. First the participant is sampling Sprite and Pepsi and asked to identify Pepsi. In this case, Sprite is the noise and Pepsi is the signal, and the two are very different. It will be easy for the participant to correctly identify Pepsi. Decisions would be expected to be made with high accuracy. Next, the participant is sampling Pepsi and Coca-Cola and asked to identify Pepsi. In this case, Coca-Cola is the noise and Pepsi is the signal. The noise and signal are very similar, and it will be more difficult for the participant to correctly identify Pepsi. More errors are expected when making these decisions. In SDT, the ability to distinguish between signal and noise is called discriminability (d').

Correct decisions also depend on the bias of the decision maker. Conclusions can be affected by biasing the decision making process. For example, consider a participant sampling

Pepsi and Coca-Cola and asked to identify Pepsi. Without any influence, he or she should not have any bias. Decisions are made based entirely upon information gathered from the stimulus. The person's bias can be shifted by providing incentives for one over the other. Say the person is rewarded $1 for each time Pepsi is correctly identified, and no punishment is made for wrong decisions. That person will shift his or her bias to select Pepsi more often. Decision bias can also be shifted by informing the participant about the relative rates of signal and noise either directly or through experience. Say the person is informed of a 3:1 ratio of Pepsi to Coca-Cola. The person will shift his or her bias to select Pepsi more often because the probability of Pepsi is higher. In SDT, the bias of a decision making process is called the criterion (*c*).

In signal detection theory, decisions fall into four categories. Correct rejection refers to correctly deciding a stimulus is just part of the noise, which correlates to failing to reject the null hypothesis when there is no true effect. A miss occurs when the decision is that a stimulus is part of the noise, but it was actually a relevant signal stimulus. This maps onto a Type 2 error, failing to identify a true effect. A hit is correctly deciding a stimulus is a relevant signal, which correlates to correctly rejecting the null hypothesis when there is a true effect. False alarms occur when deciding a noise stimulus is a relevant signal. This maps onto a Type 1 error, rejecting the null hypothesis when there is no actual effect. The greater the similarity between the stimulus and the noise, the less likely it is to make correct decisions (correct rejections or hits). Increasing difficulty in making correct decisions is not addressed by simply changing a response threshold; for example, setting a more liberal criterion will lead to more hits, but also more false alarms.

The trade-off between false alarms and hits is most easily represented by a receiving operator characteristic (ROC) curve. An ROC curve shows the efficacy of different tests based on the ratio of hits and false alarms. The area under that curve quantifies the discriminability of

the test. Figure 2 shows hypothetical ROC curves for four tests. A test that does not discriminate well will have a hit to false alarm ratio close to 1, demonstrated by the 45˚ diagonal of the ROC space, with the area under the curve only equal to 0.5. As discriminability increases, the ROC curve approaches the vertical axis, and the more accurate the test. Differing criteria are also able to be plotted on an ROC curve, as shown by the three points on the "excellent" ROC curve in Figure 2. Point 1 has a more conservative criterion while point 3 has a more liberal criterion. These points represent different biases in responding. The liberal bias indicated by point 3 will result in more true positives but will also increase false alarms. Conversely, the conservative bias indicated by point 1 will hold the false alarm rate down but will also decrease the number of true positive responses (i.e., increase the miss rate). A good test will not only have high discriminability, but will also have a fairly neutral criterion, such that the false alarm rate is contained without missing too many true positive results; this balanced perspective is indicated by point 2.

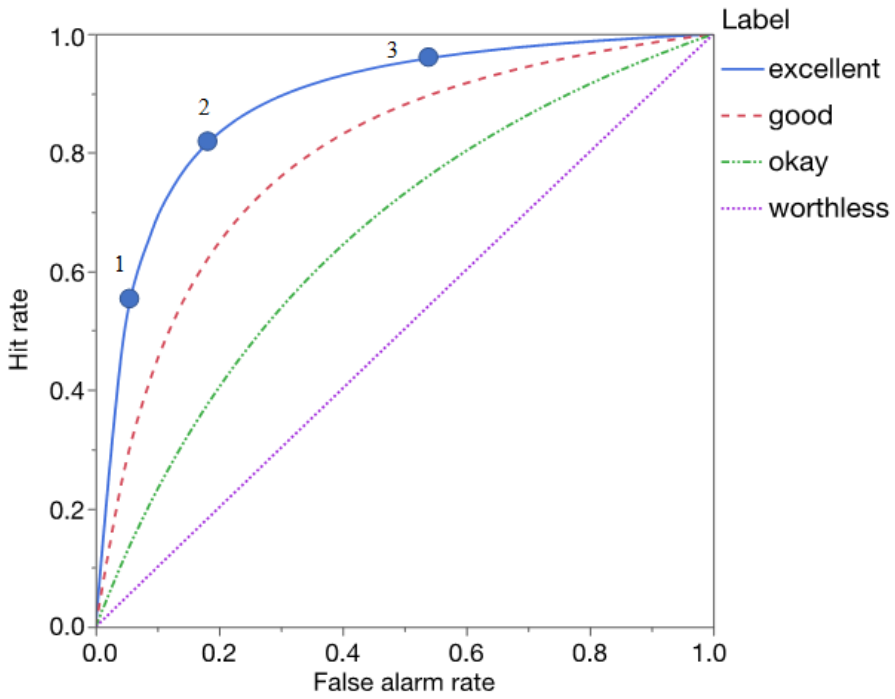

Figure 2 Receiving operator characteristic (ROC) curves. This figure shows hypothetical ROC curves and three decision points.

Signal detection theory can be used to explain the effectiveness of different analytical techniques. It can show the effect that each technique has on the discriminability between noise (differences or slopes arising due to error variance) and signal (replicable differences related to the predictor of interest). If one technique outperforms another, it could be due to an increase in discriminability. Signal detection theory also calculates decision bias and can express how each analytical technique biases the decision making process. If a technique performs particularly poorly, it could be due to a bias toward false alarms (Type I errors) or misses (Type II errors). Signal detection theory can inform on both of these situations.

When deciding the effectiveness of an analytical technique, it is important to look at hit rate and false alarm rate (MacMillan & Kaplan, 1985). The hit rate is related to the power of a test (1-Type 2 error rate). Statistically, power is the ability of the test to detect a significant result

when the groups differ. Similarly, the false alarm rate can be considered Type 1 error rate. Statistically, Type 1 error is detecting a significant result when the groups are not different.

This research is interested in determining the relationship between the different analytical techniques and their discriminability, criterion, hit rate, and false alarm rate. Understanding these relationships will help inform researchers of ideal conditions for each analytical technique and help them select the most appropriate technique for their data.

Ideal Reaction Time Models

There are a number of distributions suggested for modeling reaction time data. Among these models are the ex-Gaussian, Lognormal, Weibull, and Gamma distributions (Van Zandt, 2000). Each of these models is suggested because of its ability to model skewed data, although none of them are correct. George Box asserts that "all models are wrong but some are useful" in several publications. No model can exactly capture the nature of the population though some can provide meaningful approximations. The goal of any research is to model the data in order to provide an economical description of the population. Since all models are approximations, the researcher must focus on what is "importantly wrong" with the models, such as blatantly fitting a straight line to curvilinear data. As Box stated, "it is inappropriate to be concerned about mice when there are tigers abroad".

The ex-Gaussian is a convolution of the exponential and Gaussian distributions. The exponential and Gaussian portions model a response and decision process (Palmer et al, 2011). It is the most commonly used model for RT data, although evidence exists to suggest that it is not the generating distribution for RT data (Van Zandt, 2000). Theoretically, the Gaussian portion could produce negative values, while the exponential distribution cannot have negative values (Dolan, Van der Maas & Molenaar, 2002). RTs cannot be negative, so the ex-Gaussian

distribution cannot be the underlying distribution despite its utility in modeling many observed RT distributions and its ability to nicely model two sources of variability (exponential and Gaussian).

The lognormal distribution also fits RT data well. This distribution fits best when the log transformed distribution is normal. This occurs when the log transformed data, *ln(DV)*, is normally distributed. (Ulrich & Miller, 1998). Many RT distributions do have this property, so the lognormal distribution has been used to model RT data. However, the lognormal distribution was discovered without theoretical derivation (Dolan, Van der Maas & Molenaar, 2002). Although the lognormal distribution has been successfully fit to RT data, no one has proposed a theoretical basis connecting the parameters to cognitive processes. Not all RT data are skewed such that *ln(DV)* is normally distributed. So, this distribution would work well for special types of RT data in which that assumption is met, but it may decrease discriminability or create decision bias for situations when the distribution is not strictly lognormal. Of course, there are several ways a researcher may test the normality of transformed data. This research aims to demonstrate the importance of selecting an appropriate transformation, if that technique is to be used.

The Weibull model is commonly used in modeling survival data. Survival data occurs when the DV measurement is time until a response occurs. Morris water mazes are a type of survival (Jahn-Eimermacher, Lasarzik, & Raber, 2011). A rat is placed in a pool and must swim to a target location. The DV measured is the time until the rat reaches that location. Survival data also occurs in neuropsychopharmacology. A drug is administered and the DV measured is the time until the drug relieves symptoms, or time until the drug causes other symptoms. Reaction times are a specific type of survival data where the response is an actual physical response, such

as pressing a lever or selecting an option. Because the Weibull distribution is the gold standard of modeling survival data, it has been used to model RT data as well. This model is flexible and can fit many different levels of skew. It has been shown to model memory search RTs well, but not other types of RT data, such as visual search times (Palmer et al, 2011).

One last model used to fit RT data is the Gamma distribution. Lawless and Crowder (2004) define the probability density function following the most commonly used parameterization as

$$\frac{1}{\Gamma(\gamma)\theta\gamma} x^{\gamma-1} e^{-\frac{x}{\theta}} \quad (1)$$

where ɣ js the shape parameter and θ is the scale parameter.

This distribution is the sum of exponential distributions (Palmer et al, 2012; Lo & Andrews, 2015) such that, if a random variable $X_i \sim \exp(\theta)$, then $\Sigma X_i \sim \text{Gamma}(\gamma,\theta)$. This implies that reaction times may be the result of multiple basic processes, summing together to create the observed behavior. Each basic process (i.e. visual stimulation, initial processing, motor movements, decision making processes, etc.) is assumed to follow an exponential distribution. The final RT is the combination of all of these processes and, when summed together, culminates in a Gamma distribution (Lo & Andrews, 2015). The scale parameter, θ, is the average of all the different scales in the individual exponential processes (Palmer et al, 2011; Dolan, Van der Maas & Molenaar, 2002). The scale parameter, θ, is the most critical because covariates, or other predictor variables, enter the Gamma distribution through this parameter (Lawless & Crowder, 2004) such that:

$$\theta = \beta_0 + \beta_1 X_1. \quad (2)$$

where β_0 represents the subject intercept and β_1 represents the subject slope for predictor X_1. Of course, this equation can be extended to allow for more covariates. The subject intercept is a

constant effect; an individual baseline. The subject slope is a changing effect; how the subject interacts with the variable, X_1. While both parameters affect the scale of the Gamma distribution, β_1 has a more drastic effect than β_0. As either parameter increases, the graph of the probability distribution widens (Fig 3). The shape parameter, ɣ, simply referred to as shape, reflects the number of exponential processes contributing to the distribution. Shape is usually consistent among varying degrees of skew in RT data. This could imply the shape parameter is reflecting larger cognitive processes, such as initial processing (Palmer et al, 2011). A third parameter, the shift or location, is sometimes included. This parameter simply moves the location of the distribution along the x-axis (Palmer et al, 2011). This could occur by adding a delay to the task. The scale and shape of the distribution would not change, but RTs would uniformly be shifted in the positive direction as the shift parameter increases.

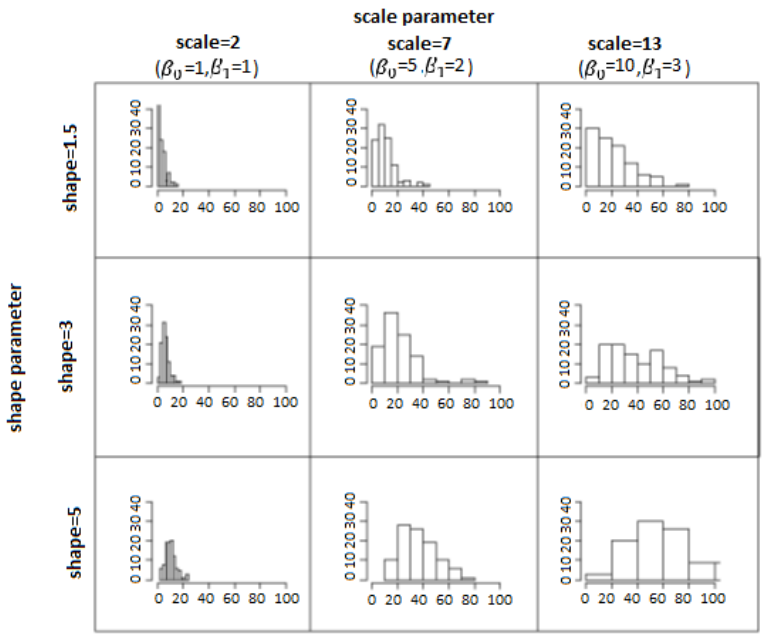

Figure 3. Randomly generated histograms. This figure shows nine histograms of randomly generated gamma distributions for a sample size of 100. Columns represent differing scale parameters, as a result of varying subject means and slopes. Rows represent differing shape parameters.

Preliminary Research

A preliminary study was conducted with only a few of the conditions of interest. Reaction time data were simulated from a Gamma distribution for 30 subjects, in two conditions, with 15 trials in each condition. This resulted in a total of 900 observations. This simulated a repeated measures, within-subjects design. A total of 100 random samples were simulated for four effect sizes. These effect sizes were achieved by changing β_1. Because β_1 was allowed to vary ($\beta_1 \sim N(\mu, 0.05)$), μ was adjusted to change the β_1 value. For no effect, μ was set to 0.00; 0.25 for a small effect; 0.50 for a medium effect; and 1.00 for a large effect. From these 100 random samples, false alarm and hit rates were calculated for three analytical techniques: ignoring distributional assumptions, aggregation, and truncation.

The results indicated relatively low power for all three analytical techniques (Table 1), particularly for small and medium effect sizes. Ignoring distributional assumptions had an acceptable false alarm rate albeit a bit conservative, 2%. However, even with a large effect size, this technique was only able to detect a significant difference 56% of the time. Aggregation also had an acceptable false alarm rate, 3%. Again, even with a large effect size, aggregating only allowed for a 59% detection rate. The most surprising results occurred when the data were truncated (±2.5 sd). Although the power was good (42% hit rate for small effects and 80% for large effects), the false alarm rate drastically increased to 37%.

Table 1

Percent of significant results from preliminary study

Technique	None	Small	Medium	Large
Ignore Assumptions	2	7	9	56
Truncation	37	42	45	80
Aggregation	3	13	12	59

From these results, *d'* and *c* were calculated (Tables 2 & 3). A criterion of 0 means there is no decision bias, where Type 1 and Type 2 errors are equally consequential. When *d'* equals 0 the two conditions are completely overlapping and the technique is unable to identify a significant result from a non-significant one. In this case the technique is unable to discriminate between the null and alternative hypotheses. As *d'* moves away from 0, discriminability increases.

Truncation appeared to reduce discriminability (*d'*) compared to the other two techniques. This effect was more apparent for smaller effect sizes. It also appeared to reduce the criterion values. The reduced ability to discriminate between the conditions and the more liberal criterion values led to the drastic increase in false alarm rate. Ignoring distributional assumptions appeared to have similar effects on *d'* and *c* as aggregation. Both techniques had low power and low false alarm rates. In the field, a significance level of 5% is standard, which corresponds to a conservative criterion value. Even with this standard, ignoring assumptions and aggregation still had an overly conservative criterion. Even with very large difference between the groups, both were only able to detect the difference about half of the time. These results suggest that when data are truncated, the analysis will generate more false alarms (Type I errors), whereas ignoring distributional assumptions and aggregation will generate more misses (Type II errors).

Table 2

Discriminability (d') values for each technique and effect size

	Effect		
Technique	Small	Medium	Large
Ignore Assumptions	0.58	0.71	2.2
Truncation	0.13	0.21	1.17
Aggregation	0.75	0.71	2.11

Table 3

Criterion (c) values for each technique. These values are assumed constant across effect sizes as it is a measure of decision bias.

	Effect		
Technique	Small	Medium	Large
Ignore Assumptions	2.05	2.05	2.05
Truncation	0.33	0.33	0.33
Aggregation	1.88	1.88	1.88

Hypotheses

General hypotheses.

Overall, when the number of observations increases, parameter estimates should be more accurate, discriminability should increase, and criterion values should not be effected. Accordingly, since increasing the number of subjects or the number of trials results in more observations, both of these should result in the preceding conclusions. Additionally, analyses that resemble the simulated data more closely (i.e. fitting a Gamma distribution) and the least amount of adjustments made to the data (i.e. aggregation, truncation) should result in better parameter recovery and discriminability.

Ignoring distributional assumptions.

When the distributional assumptions are ignored, the error variance in the distribution is inflated due to the long upper tail (Dixon, 2008). This increase is likely to result in poor discriminability and overestimation of the means. In other words, a low *d'* and a conservative *c* is expected (Tables 2 & 3). This added variability will result in larger standard error, resulting in poor parameter estimation.

As sample size increases, there will be less variability in the estimates (Miller, 1998). This decrease in variance should increase discriminability, while no change is expected in criterion, and parameter estimates should be more accurate. However, increasing sample size will not change the skew of the data. Applying a normal model to a less variable, skewed dataset will likely offset any increase in *d'* caused by the increased sample size. Therefore, a low *d'* (relative to a proper model) and conservative *c* are still expected. As effect size increases, discriminability is likely to increase. So, *d'* is expected to increase and *c* is expected to stay the same.

Aggregation.

When applied to RT data, aggregation hides the variability in the distribution and inflates Type 1 error rates (Lorch & Myers, 1990). It creates estimation bias and distorts the mean distributions (Wolfe, Palmer & Horowitz, 2010). For RT data, this bias and distortion will likely present as overestimates, since means are pulled in the direction of the longer tail of a distribution. Masking the variability and inflating Type 1 error rates suggests low discriminability, while overestimating the mean suggests a moderate to conservative criterion value. In other words, a low *d'* and a conservative *c* value would be expected (relative to a proper analysis). However, preliminary results suggest the Type 1 error rates did not increase for this technique, at least for a discrete variable with the parameter values specified in the simulation. The estimation bias and distortion of the mean distributions is likely to result in biased and distorted parameter estimates.

As sample size increases, the distribution of the means will approach normality. Increasing sample size also decreases the standard deviation, which will decrease Type 1 error and increase discriminability. This could decrease the estimation bias created by the

aggregation and counteract the inflation in Type 1 error. However, the mean is still not the appropriate measure of central tendency for RT data, so there will still be some bias present in the estimations. Therefore, increasing sample size should result in an increase in *d'* and a conservative *c* value. As effect size increases, *d'* is expected to increase and criterion values will stay the same. This pattern is also suggested by the preliminary study (Table 1).

Transformations.

Transformations are meant to reestablish normality (Baayen & Milin, 2010). However, the current study applies a logarithmic transformation to data that follows a Gamma distribution. While the log transformation helps produce normality (Van Breukelen, 2005), it is most appropriate when *ln(DV)* follows a normal distribution (Ulrich & Miller, 1994). Log transforming the DV should increase normality and decrease the error variability producing a moderate to high *d'* and an unbiased criterion value. The transformed data should have reasonable parameter estimates, since the transformation is attempting to produce normality to the distribution. However, applying the log transformation will not completely produce normality, since the simulated data follows a Gamma distribution. Therefore, the parameters will likely be better estimated by the Gamma distribution.

As sample size increases, error variability should decrease thereby increasing discriminability. However, a log transform is still being applied to data following a Gamma distribution, so some of that increase will be offset by applying the wrong transformation. Therefore, a slight increase in *d'* and no change to the criterion is expected. Increasing the sample size will likely lead to more accurate parameter

estimates, although the standard errors will still not be ideal since the wrong transformation is being applied. Increasing the effect size increases the differences between conditions for the discrete variable and increases the slope for the continuous variable, but does not affect the criterion or standard error. Consequently, *d'* is expected to increase, while *c* and parameter recovery will stay the same.

Truncation.

Truncation introduces estimation bias and reduces power (Ulrich & Miller, 1994). Eliminating potentially influential observations will likely decrease discriminability. However, the truncation will reduce the variability of the distribution. Reducing variability would likely increase power, so *d'* could increase. Because more observations in the upper tail get eliminated than lower values, parameters will likely be underestimated. This estimation bias will likely result in a lower *d'* and liberal criterion are expected. Although preliminary results suggest a slightly conservative criterion, it is much more liberal than is expected for a 5% significance level.

As sample size increases, standard error decreases. This decrease in variability would increase power and could lead to an increase in discriminability. Therefore, a slight increase in *d'* and no change to the criterion is expected. While larger samples will lead to reduced error variability, they will not decrease the skew of the data. When truncating according to number of standard deviations from the mean, a similar proportion of data will be eliminated, regardless of sample size. Accordingly, accuracy in parameter estimates are not expected to increase nor decrease. Increasing the effect size increases the differences between conditions for the discrete variable and increases the slope for the continuous variable, but does not affect the estimation bias or standard error.

Consequently, *d'* is expected to increase, while *c* and parameter recovery will stay the same.

Generalized model fitting.

Fitting a generalized model, the Gamma distribution, to the raw data is expected to increase *d'* and, in order to maintain the standard significance level, set a conservative criterion value. When the data are generated from a Gamma distribution and a Gamma model is fit to it, the power and parameter recovery should be maximized and the Type 1 error minimized.

While smaller sample sizes will introduce more variability, fitting the appropriate model to the data should still produce the highest power and lowest Type 1 error of all the techniques. Similarly, the Gamma fit should outperform the other techniques, regardless of effect size. Accuracy of the parameter estimates should increase as the sample size increases.

Chapter 2- Methods

Participants

The current research simulates data for varying sample sizes and characteristics. Since this study involves simulating data, instead of collecting it, specific participant information is not available or relevant. Because reaction time (RT) data are used in many different types of experiments, it is important to consider a range of sample sizes. Sample sizes (the number of participants) were chosen to include the range of the most common values many of those found in RT datasets in psychology: n = 30, 60, and 120. Although some RT research may have fewer than 30 participants, it is more common to have at least 30. Including smaller samples would introduce a whole new set of factors,

such as higher variability in the means and bias in the parameter estimates (Miller, 1998). Small samples can require different analytical techniques than those investigated in the current research (Van Zandt, 2000) or the collection of much more data per participant. It is the intent of the current research to investigate the general effects of common analytical techniques on RT data.

Materials

Data are simulated using the statistical software package SAS® 9.4 (Appendix A). Analyses are conducted using JMP Pro 12. Probit regression analysis are conducted using R.

Procedure

To assess the range of effects that different analytical techniques have on RT data, the current research employs Monte Carlo simulation combined with signal detection theory to evaluate model efficacy.

Defining the Parameters

The current research utilizes the two-parameter Gamma distribution to simulate RT data. A location parameter was not necessary because the lower bound of reaction times (RTs) is 0. Because the shape parameter is fairly consistent for RT data (Palmer et al, 2012), it is fixed in simulating the data (but not in the analysis). Based on observed parameter estimates calculated from several psychology datasets involving RT data, it was decided, empirically, to fix the shape parameter at 1.5. Even though these datasets varied in task, species evaluated, and average RT, the shape parameter was consistently about 1.5. The scale parameter, however, varied systematically. Predictors enter the

Gamma distribution through its scale parameter such that $\theta = \beta_1 x + \beta_0$. β_0 and β_1 values are simulated to create small, medium, and large effect sizes using a within-subject manipulation; β_0 models the baseline RT whereas β_1 models the effect of the independent variable relative to this baseline. The effect sizes are determined by calculating Cohen's *f* (Cohen, 1992). β_0 was decided, empirically, to be set at 6.8. Several datasets were analyzed involving RTs of differing lengths and this value was consistently about 6.8. However, the simulation allows for random, normal variation around β_0, such that $\beta_0 \sim N(6.8, 1)$. β_1 is also allowed to vary, $\beta_1 \sim N(\mu, \sigma)$. For a continuous predictor, the μ and σ parameters were set to 0.00, 0.06 for no effect; 0.25, 0.06 for a small effect; 0.50, 0.11 for a medium effect; and 1.00, 0.22 for a large effect. For a discrete predictor, the μ and σ parameters were set to 0.000, 0.024 for no effect; 0.100, 0.024 for a small effect; 0.200, 0.044 for a medium effect; and 0.400, 0.088 for a large effect.

Simulation Details

The current research focuses on either a discrete two-level or a single continuous predictor variable. Beginning with a single predictor (discrete or continuous) will establish a baseline relationship between the analytical techniques and RT data. Because the covariates enter into the scale parameter in an additive manner, conclusions about multiple predictors are able to be drawn but the extension of these findings to more complex designs will await further research.

The parameter β_1 varied to produce different effect sizes. For the discrete predictor, β_1 *=0, .1, .2, .4*, and the predictor, *x*, took on the values of 1 and 5. This designation was chosen to produce a typical difference between the group means when

multiplied by the slope parameter. For the continuous predictor, β_1 = *0, .25, .50, 1.00*, and the predictor, *x*, was sampled at unit intervals from 1 to 10 so that $\theta = \beta_1\left(1+10\left(\frac{x-1}{9}\right)\right)+\beta_0$. The particulars of this equation were arbitrary but ensured equally spaced sampling of the continuum. Slope, β_1, values were decided using Cohen's *f* to create small, medium, and large effect sizes. In both cases, when β_1 = *0*, there was no effect. From this, false alarm rates could be found. These parameters were used to compare group effects.

Data are simulated to create a completely within-subjects dataset with a varying number of subjects (30, 60, and 120), a discrete predictor with two levels or a continuous predictor, and 5 or 10 trials per subject. These numbers of trials are chosen to simulate realistic RT data environments. To create a sampling distribution, 1,000 samples are simulated with a full factorial combination of predictor type (discrete vs. continuous), number of subjects (30, 60, or 120), number of trials per subject (5 or 10), and effect size (0.00, 0.25, 0.50, and 1.00 for continuous predictor; 0.0, 0.1, 0.2, 0.4 for discrete predictor).

Once the data were generated, several analytical techniques were used for analyses. These analyses produced estimates of the average intercept and the average difference (for a discrete predictor) or slope (for a continuous predictor). In other words, I investigated the mean effect under the control group (difference or slope of 0) and the mean effect under the treatment group (difference or slope not equal to 0). These estimates were compared to the true means to compute parameter bias for both the

intercept and the difference or slope. Finding the true means depended on which technique was being used.

Determining Parameter Bias

A generalized mixed model was conducted on the simulated model using PROC GLIMMIX in SAS ©. Analyses of each technique provided estimates of the intercept and difference or slope parameters. For clarity, b_0 and b_1 are defined as the estimates of the intercept and difference/slope parameters, respectively, when the simulated data are analyzed using the GLIMMIX normal model (for ignoring assumptions, aggregation and truncation); l_0 and l_1 are the estimates of the intercept and difference/slope parameters when the log of the simulated data are analyzed using the GLIMMIX normal model (for transformation); and g_0 and g_1 are the parameter estimates when the GLIMMIX Gamma model is used (for applying Gamma).

Generally, bias was computed by finding the difference between the population parameter and the estimated value from the analysis. This was dependent on the technique used. When a discrete predictor was used, each analysis estimated an average intercept value and an average difference of means. When a continuous predictor was used, the analyses estimated the average intercept value and an average slope. Some calculations involved arithmetic means while others involved geometric means. The techniques of ignoring distributional assumptions, aggregation, truncation and applying the Gamma model all involved the arithmetic mean. Applying a log transformation involved the geometric mean. Accordingly, bias was calculated by comparing the estimates to the appropriate population mean for each technique.

When a discrete predictor was used, calculations were straight forward. The arithmetic population intercept mean was determined by $(6.8+5\beta_1)$*1.5 and the arithmetic population mean difference was determined by $-(4\beta_1$*1.5$)$. The geometric population intercept mean was determined by $(6.8+5\beta_1)$ and the geometric population mean difference was determined by $-(4\beta_1)$.

When a continuous predictor was used, calculations were not as clear-cut, particularly for the log-scale techniques. The arithmetic population intercept mean for the aggregated, truncated, and raw techniques was determined by $\left(6.8+\left(1-\frac{10}{9}\right)\beta_1\right)$*1.5 and the arithmetic population slope mean was determined by $\left(\frac{10\beta_1}{6}\right)$.

When log-scale techniques were used, deteriming theoretical means was tricky because they involve fitting a straight line through a curve and vice-versa. For the log transformed data, the original linear relationship becomes curved in the log transformed space so applying the normal model would involve fitting a straight line through a curve. For the Gamma model, this means fitting a curve to a straight line. Due to the model misspecification, the parameter bias for this case was determined by comparing the estimated intercept and slope to the best fitting straight line through the curve.

Lastly, when computing intercept bias, b_0 was compared to the population arithmetic mean, l_0 was compared to the population geometric mean, and $\exp(g_0)$ was compared to the population arithmetic mean. When computing *difference* bias, b_1 was compared to the population arithemetic mean, $(e^{l_0+l_1}-e^{l_0})$ was compared to the

population geometric mean, and $\exp(g_1)$ was compared to the population arithmetic mean. When computing *slope* bias, b_1 was compared to the population arithmetic mean, and l_1 and g_1 were compared to the population slope means derived from the best fitting line.

The absolute values of these biases were used to compare accuracies. Since these values were not normally distributed, the fourth root transformation was applied so that a full factorial ANOVA could be used to determine effects of the number of trials per subject, sample size and effect size. Both estimation bias and accuracy were determined for each run to be used in the ANOVA. In this manner an average estimation bias and average accuracy for each trial size, effect size and sample size combination could be found.

This research subjected the simulated datasets to multilevel modeling using five analytical approaches: ignoring distributional assumptions; aggregation; transformation; truncation; and generalized multilevel modeling using the Gamma distribution. To simulate ignoring the distributional assumptions, a general multilevel linear model is fit to the raw data. No transformations or adjustments are made to account for the skew. To simulate aggregation, raw data are averaged across trials. For the discrete model, this will result in an average score per participant, per condition. For the continuous model, averaging across trials would eliminate the continuity of the data, resulting in only an average score per participant; all within-participant variation would be lost. Accordingly, for this research, averaging across trials will only occur for the discrete model, and a general linear model is then fit to these aggregated values. To simulate transformation, the commonly used log-transform is applied to the raw data and a general multilevel

linear model is fit to these transformed values. To simulate truncation, this research uses a cutoff of ± 2.5 standard deviations from the mean. A general multilevel linear model is then fit to these truncated values. Lastly, to simulate the recommended fitting of a Gamma distribution to data that were sampled from a Gamma, a generalized multilevel linear model, specifying the Gamma distribution, is fit to the raw data (See Appendix A).

These parameters were chosen to reflect common situations involving RT data. The specific techniques are some of the most commonly used techniques in the field (Baayen & Milin, 2010). A quick survey of a well-respected psychological journal revealed that 38% of the articles analyzed RT data. Of these particular articles, 82% used aggregated data, 55% used some type of truncation based on standard deviations from the mean, and 18% used a transformation. Researchers may be unfamiliar with multilevel modeling and choose to ignore the skew in RT data, oftentimes believing the linear-scale model to be robust to fluctuations from normality. It is for this reason that this research investigates the effects of ignoring distributional assumptions. Additionally, many times the general trends are typically of interest so researchers aggregate across trials or across subjects. Since the distribution of means is normally distributed, the CLT is often the justification for using aggregated data. When the skew in the data is accounted for, several transformations may be used for analyzing RT data (Baayen & Milin, 2010; Dolan, Van der Maas, Molenaar, 2002; Palmer et al., 2011; Ratcliff & Murdock, 1976; Van Zandt, 2000). This research specifically investigates the effects of the log transformation because it is the most common in the literature. Another common linear-scale technique used to address the skew in the data is truncation. Most of the time the truncation is based on the mean so this research investigates the effects of that type of

truncation. Lastly, this research is interested in discovering the benefits, if any, of determining the underlying distribution and using that model. Since the data are generated from a Gamma distribution, this model is used in the analyses.

Signal Detection Theory

Once each of the data sets has been simulated and analyses performed, an overall evaluation of each technique can be considered. When a technique discovers a significant result when μ (the mean of the subject slope, β_1) = 0 (i.e., conditions are not different), it is counted as a false alarm. For the discrete model, this would mean finding a difference between conditions when the conditions are not different. For the continuous model, this would mean claiming a non-zero slope when, in fact, it does equal zero. This will determine the Type 1 error rate of that technique. When a technique discovers a significant result when $\mu \neq 0$ it is counted as a hit. For the discrete model, this means finding a difference between conditions when they are, in fact, different. For the continuous model, this means finding a non-zero slope when it is truly non-zero. These two factors determine the power of that technique. From the hit rate and false alarm rate, the discriminability (d') and criterion (c) measures from signal detection theory (SDT) can be calculated. When the criterion is not neutral, for a given significance level, the decision making process is said to be biased. Accordingly, the criterion value is often referred to as a measure of bias. Discriminability and criterion values are generally defined using the following equations from MacMillan & Creelman (1991):

$$d' = z(\text{Hit Rate}) - z(\text{False Alarm Rate}), \text{ and} \tag{3}$$

$$c = \frac{z(\text{Hit Rate}) + z(\text{False Alarm Rate})}{-2}. \tag{4}$$

The current project uses probit regression to estimate discriminability and criterion (DeCarlo, 1998; Wright & London, 2009). Two variables were created to determine the specificity and power of each test. The predictor variable classified each result as .5 when a difference was coded in the simulation (i.e. when the effect size was not equal to zero) and -.5 when there was not a difference (i.e. when the effect size was zero). The outcome variable recorded whether the test had determined significance; 0 for a non-significant result and 1 for a significant result. These two were entered into a generalized linear model specifying a probit regression to extract *d'* and *c* values. A full-factorial probit including the number of trials per subject, analytic technique, and sample size as moderators was conducted to determine differences between techniques. The intercept estimate corresponds to the criterion as it is a baseline measurement of decision making, since techniques will produce different biases when determining significance. Then, the slope estimate measures discriminability, as it determines differences in the distributions.

Analyzing SDT parameters gives a better description than traditional goodness of fit (GOF) measures. Goodness of fit measures help determine how well a model fits the data. However, they do not measure flexibility or generalizability of the models (Pitt & Myung, 2002). It is important that any analytical technique estimates the parameters and shape of the distribution well, but sometimes the model with the best fit is simply the model with the most flexibility (Cutting, 2000). A model that is too flexible will not generalize to other datasets and does not summarize the cognitive processes well enough (Veksler, Meyers & Gluck, 2015).

Parameter Estimation

While analyzing SDT parameters provides a general idea of how well the different analytical techniques model the simulated data, it is also important to look at how well each technique recovers the actual value of the parameters. Since the data has been simulated, we know the true value of each parameter in the model. Each analytical technique will also provide estimations of those parameters based on maximum likelihoods. To give an idea of the ability of each condition to estimate the parameters, the estimation errors will be compared. To measure parameter bias, estimates will be directly compared to their theoretical value. To measure accuracy, the fourth root of the absolute value of the errors will be used.

This research will combine the general model fitting ability from the SDT parameters with its ability to recover the actual parameter values to determine the best conditions for using each analytical technique. The best model is one which provides a good fit of the model while maintaining high accuracy in parameter estimation. In the data simulation, a linear relationship between the number of trials and mean RT was produced. In retrospect, this relationship does not mimic the relationship found in actual RT data.

Chapter 3- Results

Signal Detection Theory Analyses

The benefit of Monte Carlo simulations is being able to know the reality of the data. When the decision and the reality are both known, the number of hits, correct rejections, false alarms, and misses were able to be determined (Table 4).

Table 4

Percent of total results for each technique

			Aggregation		Gamma		Ignoring		Transform		Truncation	
			In Reality									
			Not Sig	Sig	Not Sig	Sig	Not Sig	Sig	Not Sig	Sig	Not Sig	Sig
Discrete Predictors	*Decision*	Not Sig	24	36	24	35	24	36	24	40	23	35
		Sig	1	39	1	40	1	39	1	35	2	40
Continuous Predictors		Not Sig			28	22	24	19	24	23	24	20
		Sig			2	68	1	56	1	52	1	55

Key		
	Correct Rejection	Miss
	False Alarm	Hit

Since these measures are standardized, the techniques can be directly compared. Post-hoc comparisons were made for significant effects using Tukey's Honestly Significant Difference. Because technique differences were the focus of this dissertation, graphs and discussion of the main effects and interactions that do not involve technique have been relegated to Appendix B.

The hypotheses in this research can be classified into two overarching predictions:

H1. Each technique will affect discriminability and criterion differently, but applying the Gamma distribution should outperform the other techniques.

H1a. Ignoring distributional assumptions, aggregating, and truncating the data will result in low *d'* values, while transforming the data and applying a Gamma distribution will result in high *d'* values.

H1b. Both ignoring distributional assumptions and applying a Gamma distribution will result in a neutral *c* value, while aggregating and truncating the data will result in a liberal *c* value. Transforming the data should retain a neutral c value.

Table 5

Main effect results from the 4-way ANOVA

		d'			*c*		
	Predictor	DF	F	p-value	DF	F	p-value
Continuous Predictors	Trials	1	464.6	<.001	1	581.2	<.001
	Technique	3	53.3	<.001	3	75.7	<.001
	Sample.Size	2	263.4	<.001	2	261.0	<.001
Discrete Predictors	Trials	1	65.5	<.001	1	36.7	<.001
	Technique	4	28.9	<.001	4	63.2	<.001
	Sample.Size	2	132.5	<.001	2	167.8	<.001

Table 6

Two and three way interaction effects from the 4-way ANOVA

		d'			*c*		
	Predictor	DF	F	p-value	DF	F	p-value
Continuous Predictors	Technique* Sample Size	6	2.0	0.923	6	3.6	0.726
	Technique*Trials	3	5.4	0.142	3	4.8	0.190
	Sample Size*Trials	2	19.5	<.001	2	16.6	<.001
	Technique*Sample Size*Trials	6	1.5	0.960	6	7.0	0.317
Discrete Predictors	Technique* Sample Size	8	0.7	>0.999	8	2.5	0.963
	Technique*Trials	4	12.8	0.012	4	21.8	<.001
	Sample Size*Trials	2	2.5	0.253	2	1.2	0.539
	Technique*Sample Size*Trials	8	1.4	0.994	8	0.9	0.996

Effects of different techniques.

There was a significant effect of analytic technique on both the discriminability and the criterion value for discrete and continuous predictors (Table 5) and an interaction effect involving technique (technique*trials) for discrete predictors (Table 6). In order to directly compare the effects of truncation with the other techniques, discriminability and criterion values are compared only at 10 trials, as truncation did not occur when there were only 5 trials per subject. With only 5 trials, the standard deviation was large so that there were no observations that fell outside the 2 standard deviation range. Transforming the data significantly lowered the discriminability and produced a more conservative criterion for both types of predictors (Figs 4 and 5; Table 7). These results were in opposition to the original hypothesis that transforming the data would have high discriminability and a criterion that maintains a 5% significance level without creating further decision making bias in either direction.

To maintain the significance level at 5%, the criterion value will necessarily shift as discriminability changes. From equations (3) and (4), it can be derived that

$$c = \frac{d' + 2(z(False\ Alarm))}{-2} \tag{5}$$

When the false alarm rate is maintained at 5%, the relationship between *c* and *d'* becomes

$$c = \frac{d'}{-2} + 1.645 \tag{6}$$

Criterion values of zero occur when the Type 1 and Type 2 errors are equally consequential, but in the field Type 1 errors are generally controlled at 5%. Accordingly, research in the field generally utilizes a more conservative criterion to preserve an acceptable Type 1 error rate. For reference, at a 5% significance level, a *d'*=2

corresponds to a criterion value of 0.64. Anything above *c*=0.64 would be considered conservative while anything below *c*=0.64 would be considered liberal for an alpha of .05.

Table 7

Average d', c and neutral c for each technique

	Discrete			Continuous		
Technique	*d'*	*c*	*Neutral c*	*d'*	*c*	*Neutral c*
Applying Gamma	2.32	.51	.49	3.84	-.40	-.28
Ignoring Assumptions	2.18	.60	.56	3.86	-.36	-.29
Transformation	1.99	.64	.65	3.53	-.20	-.12
Truncation	2.09	.40	.60	3.70	-.29	-.21
Aggregation	2.27	.55	.51			

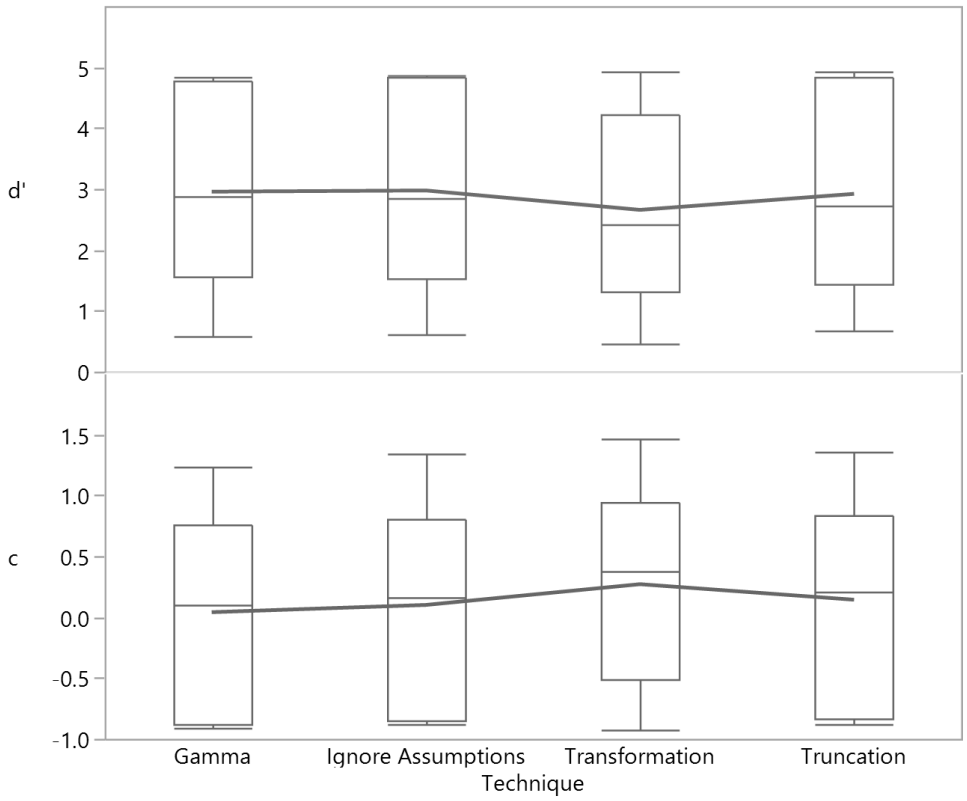

Figure 4 Discriminability and criterion for different analytic techniques for continuous predictors. Main effect plot for technique on *d'* and *c* for continuous predictors.

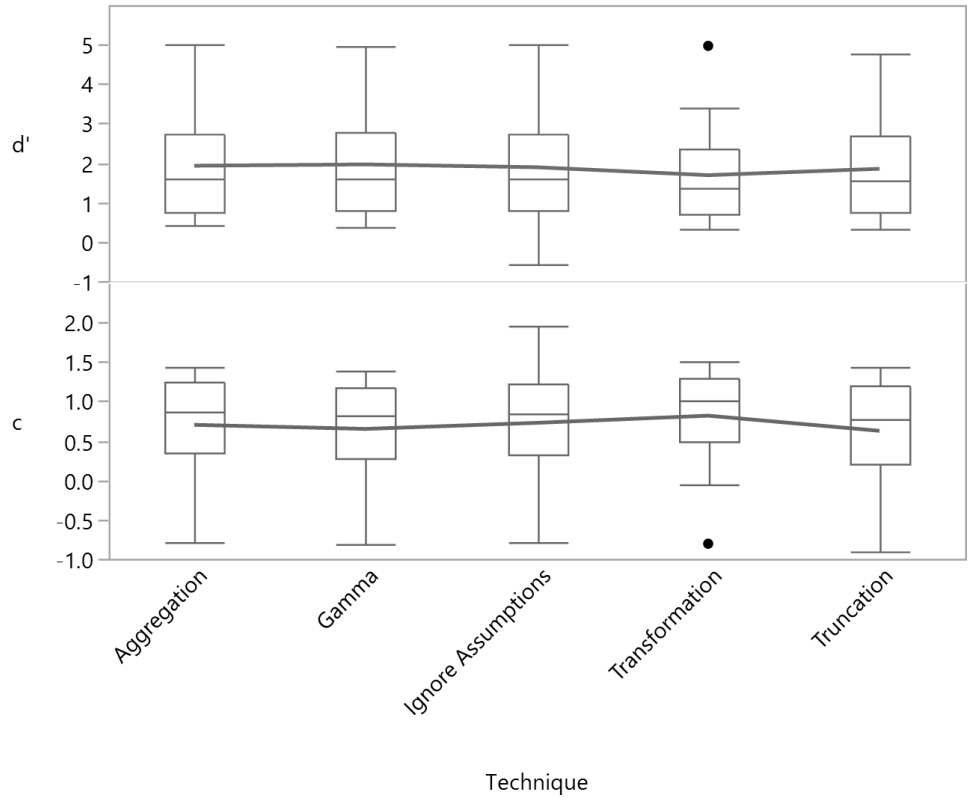

Figure 5 Discriminability and criterion for different analytic techniques for discrete predictors. Main effect plot for technique on *d'* and *c* for discrete predictors.

Comparative analyses were performed for each level of trials per subject (i.e. 5 vs 10 trials) for several reasons. This variable determines whether truncation occurs (at 10 trials) or does not occur (at 5 trials). Furthermore, the number of trials per subject is a significant factor for both types of predictors and significantly interacts with technique for discrete predictors (Table 6). It only makes sense to compare techniques within each level of trials per subject.

At 5 trials, when truncation did not occur, all techniques had similar discriminability when discrete predictors were used. At 10 trials, when truncation did occur, transforming and truncating the data had significantly less discriminability than the other techniques, which were not significantly different than each other (Table 8).

Table 8

Pairwise comparisons of d' for techniques when discrete predictors are used

	5 Trials		10 Trials	
	Estimate	p-value	Estimate	p-value
Agg-Gamma	-0.01	>.999	-0.02	0.996
Agg-Ignore	-0.03	0.989	-0.02	0.995
Agg-Log	0.10	0.386	0.21	0.003
Agg-Trunc	-0.03	0.989	0.19	0.006
Gamma-Ignore	-0.02	0.996	0.00	>.999
Gamma-Log	0.11	0.299	0.23	0.001
Gamma-Trunc	-0.02	0.996	0.21	0.001
Ignore-Log	0.13	0.155	0.23	0.001
Ignore-Trunc	0.00	>.999	0.21	0.001
Log-Trunc	-0.13	0.155	-0.02	0.995

When discrete predictors were used, at 5 trials, transforming the data had a significantly more conservative criterion than the other techniques, which were not significantly different from each other. At 10 trials, truncation was the most liberal technique while transformation was the most conservative. The remaining techniques were not significantly different from each other (Table 9).

Table 9

Pairwise comparisons of c for techniques when discrete predictors are used

	5 Trials		10 Trials	
	Estimate	p-value	Estimate	p-value
Agg-Gamma	0.05	0.133	0.04	0.189
Agg-Ignore	0.01	0.999	0.02	0.848
Agg-Log	-0.12	<.001	-0.11	<.001
Agg-Trunc	0.01	0.999	0.10	<.001
Gamma-Ignore	-0.04	0.232	-0.02	0.777
Gamma-Log	-0.17	<.001	-0.15	<.001
Gamma-Trunc	-0.04	0.232	0.06	0.024
Ignore-Log	-0.13	<.001	-0.12	<.001
Ignore-Trunc	0.00	>.999	0.08	<0.001
Log-Trunc	0.13	<.001	0.21	<.001

When continuous predictors were used, at both 5 and 10 trials, transforming the data had significantly less discriminability than the other techniques, which were not significantly different from one another (Table 10).

Table 10

Pairwise comparisons of d' for techniques when continuous predictors are used

	5 Trials		10 Trials	
	Estimate	p-value	Estimate	p-value
Gamma-Ignore	-0.01	>0.999	0.00	>0.999
Gamma-Log	0.20	0.002	0.33	<.001
Gamma-Trunc	-0.01	0.998	0.15	0.091
Ignore-Log	0.21	0.001	0.33	<.001
Ignore-Trunc	0.00	>.999	0.16	0.079
Log-Trunc	-0.21	0.001	-0.18	0.020

Also when continuous predictors were used, at 5 trials, transforming the data had the most conservative criterion while the other techniques were not significantly different from one another. At 10 trials, when truncation occurred, transforming the data was significantly more conservative than truncation, which was more conservative than both applying the gamma distribution and ignoring distributional assumptions (Table 11).

Table 11

Pairwise comparisons of c for techniques when continuous predictors are used

	5 Trials		10 Trials	
	Estimate	p-value	Estimate	p-value
Gamma-Ignore	-0.05	0.089	-0.05	0.540
Gamma-Log	-0.19	<.001	-0.26	<.001
Gamma-Trunc	-0.05	0.076	-0.15	<.001
Ignore-Log	-0.15	<.001	-0.25	<.001
Ignore-Trunc	0.00	>.999	-0.11	0.005
Log-Trunc	0.15	<.001	0.14	<.001

Parameter Recovery

With simulation studies, one may assess the effects of particular variables or techniques on parameter recovery. Since the actual parameter values are known, it is possible to quantify the amount of error in a variable or technique's estimate of that parameter. This research focuses on the estimation bias and accuracy of the parameter estimates. Here, bias differs from its definition in signal detection theory and refers to whether the variable or technique tends to over or under-estimate the actual parameter. Accuracy refers to the magnitude of the estimation bias. Larger errors indicate less accurate parameter estimates. Each of these measurements are analyzed for both the intercept and difference parameters. Tukey's Honestly Significant Difference test was used when post-hoc comparisons were necessary.

All predictors (number of trials per subject, sample size, and effect size) were centered so the intercept effect could be directly interpreted. Absolute error estimates were naturally right-skewed and a fourth root transformation was applied to produce normality. Some techniques, such as transforming the data and applying the Gamma distribution, estimate the parameters on a different scale. In this case, comparing estimates to those originating in a linear space, such as ignoring distributional assumptions, aggregation, and truncation, would not be appropriate. Therefore, the results will be discussed in two parts: techniques estimating on a linear scale (ignoring assumptions, aggregation, truncation) and techniques estimating on log scales (applying a Gamma distribution and transforming the data). Accordingly, every analysis was separated into these categories. They were further divided into two additional categories: discrete and continuous predictors. Each model uses an overall mean (*OM*) and experimental error (ε), but differ in the variables of interest.

Several models were used to assess the effects of the different linear-scale analytic techniques. To begin, four models (Table 12) were used to first determine if there were any differences among them:

Table 12

Models used for assessing significant effects of estimation bias and accuracy. Each model is comprised of an overall mean effect, technique effect, and experimental error.

Model	Assesses
$Raw\ Errors\ of\ the\ Intercept = OM + technique_i + \varepsilon_{ij}$	Bias in the intercept estimates
$Raw\ Errors\ of\ the\ Difference = OM + technique_i + \varepsilon_{ij}$	Bias in the difference (for discrete predictors) or slope (for continuous predictors) estimate
$\sqrt[4]{\lvert Errors\ of\ the\ Intercept \rvert} = OM + technique_i + \varepsilon_{ij}$	Accuracy of the intercept estimates
$\sqrt[4]{\lvert Errors\ of\ the\ Difference \rvert} = OM + technique_i + \varepsilon_{ij}$	Accuracy of difference (for discrete predictors) or slope (for continuous predictors) estimates

When discrete linear-scale predictors were used, the omnibus tests indicate significant models for all but the difference estimation bias. When continuous linear-scale predictors were used, the omnibus tests indicate significant models for all four response variables (Table 13).

The following discussion will cover the main effect of technique and interactions involving technique. The other effects are presented in the Appendix.

Table 13

Omnibus test results for all four models

	Intercept Bias		Difference Bias		Intercept Accuracy		Difference Accuracy	
	DF	F p-value	DF	F p-value	DF	F p-value	DF	F p-value
Discrete Linear Predictors	71	42.37 <0.001	71	1.20 .0124	71	140.71 <0.001	71	165.19 <0.001
Continuous Linear-scale Predictors	47	15.70 <0.001	47	79.83 <0.001	47	164.86 <0.001	47	406.65 <0.001

Intercept parameter bias.

Linear scale.

When discrete linear-scale predictors were used, there was a significant difference among the techniques, (Table 14). Ignoring distributional assumptions (*M=0.007, SD=0.56*) was not significantly different from aggregation (*M=0.005, SD=0.56*), but truncating the data (*M=-0.138, SD=0.59*) was significantly more biased than both of them (Fig 6).

Table 14

Effects model for the intercept parameter estimate

	Discrete Linear			Continuous Linear		
Source	DF	F	p-value	DF	F	p-value
technique	2	556.09	<.0001	1	7.62	0.0058
trials	1	734.00	<.0001	1	145.84	<.0001
technique*trials	2	564.17	<.0001	1	7.62	0.0058
sample.size	2	2.99	0.0505	2	1.54	0.2135
technique*sample.size	4	0.01	0.9999	2	0.04	0.9564
trials*sample.size	2	2.58	0.0757	2	6.21	0.002
technique*trials*sample.size	4	0.01	0.9998	2	0.04	0.9564
effect.size	3	2.94	0.0317	3	8.90	<.0001
technique*effect.size	6	1.36	0.2261	3	48.64	<.0001
trials*effect.size	3	1.34	0.2586	3	74.02	<.0001
technique*trials*effect.size	6	1.58	0.1472	3	48.64	<.0001
sample.size*effect.size	6	2.82	0.0095	6	0.86	0.5266
technique*sample.size*effect.size	12	0.02	1	6	0.02	1
trials*sample.size*effect.size	6	3.10	0.0049	6	2.63	0.015
technique*trials*sample.size*effect.size	12	0.03	1	6	0.02	1

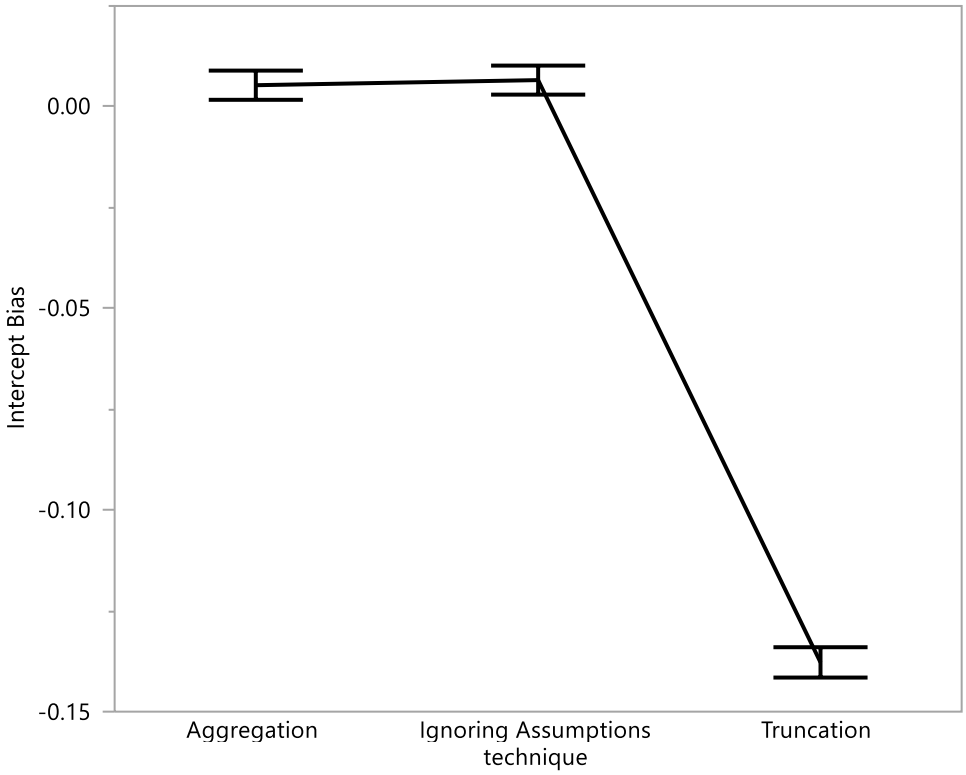

Figure 6 Intercept bias by technique. Error in the intercept parameter estimates for each discrete linear-scale technique: aggregation, ignoring assumptions, and truncation. Each error bar is constructed using 1 standard error from the mean.

Truncation greatly underestimated the intercept parameter when discrete linear-scale predictors were used while ignoring assumptions and aggregation were unbiased (Table 15). Furthermore, technique interacted with the number of trials per subject (Table 14). For 5 trials, truncation did not occur and all techniques performed similarly, but for 10 trials, truncation occurred and truncating the data caused notable underestimation (Fig 7).

Table 15

Post hoc means comparisons for technique for discrete predictors. Techniques with the same letter are not significantly different from each other.

Technique			Least Sq Mean
Ignoring Assumptions	A		0.007
Aggregation	A		0.005
Truncation		B	-0.138

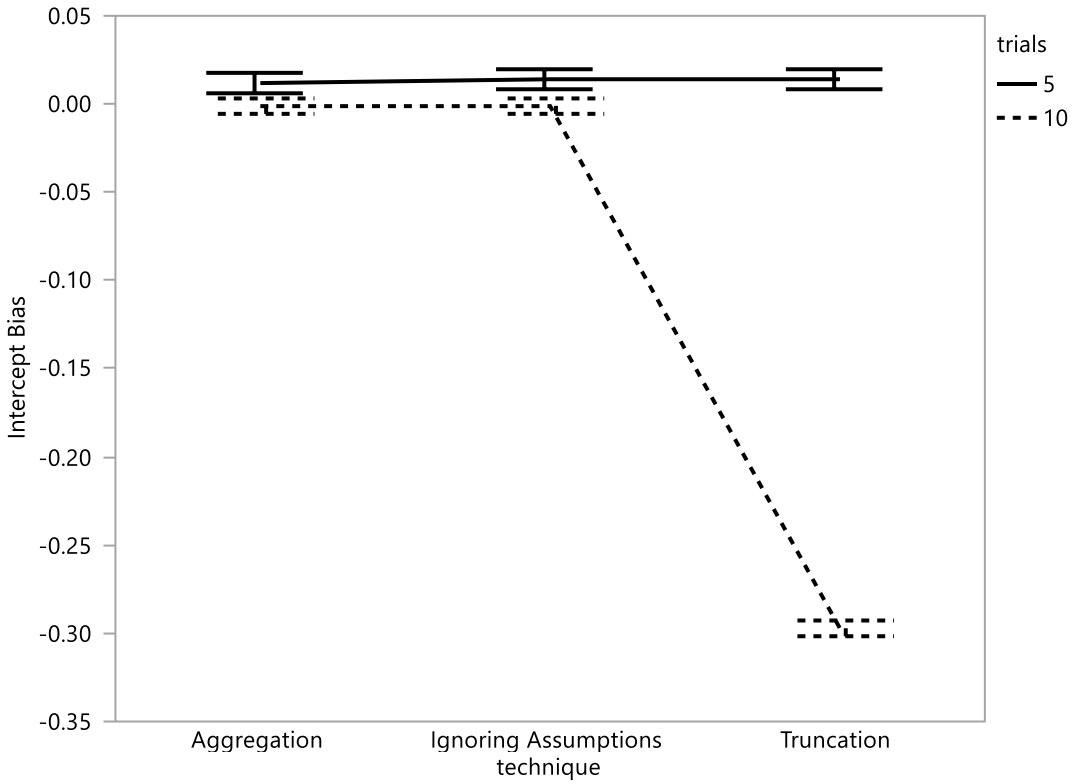

Figure 7 Bias in the intercept parameter estimates by number of trials per subject. Error in the intercept parameter estimates for each technique by number of trials per subject using discrete linear-scale predictors. Each error bar is constructed using 1 standard error from the mean.

When measuring the bias in the intercept parameter using continuous linear-scale predictors, there was a significant difference between techniques (Table 14). Ignoring

distributional assumptions (*M=-0.058, SD=1.24*) underestimated the intercept parameter more than truncation (*M=-0.027, SD=1.23*) (Fig 8) but both were biased. When the distributional assumptions were ignored, the effect that the long right tail has on the mean is being ignored, causing underestimation in the intercept parameter estimates. This is likely more detrimental than truncation because the simulated data was not too heavily skewed, minimizing the effects of the truncation.

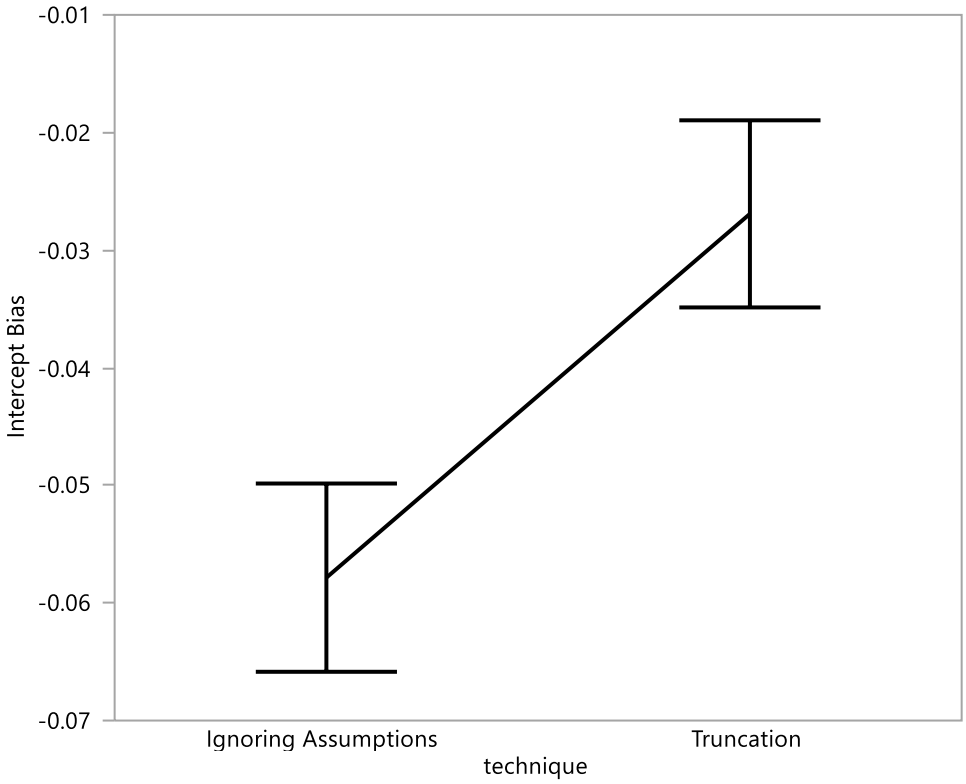

Figure 8 Intercept bias by technique. Error in the intercept parameter estimates for each continuous linear-scale technique: ignoring assumptions and truncation. Each error bar is constructed using 1 standard error from the mean.

Technique also interacted with trials and effect size (Table 14). For 5 trials, when truncation did not occur, the results were the same, but for 10 trials, the effect size caused differences between techniques. Ignoring distributional assumptions had little to no estimation

bias for all effect sizes, but truncating the data caused underestimation for small effect sizes and overestimation for larger effect sizes (Fig 9). Both of these effects were present in the significant three-way interaction (Table 14).

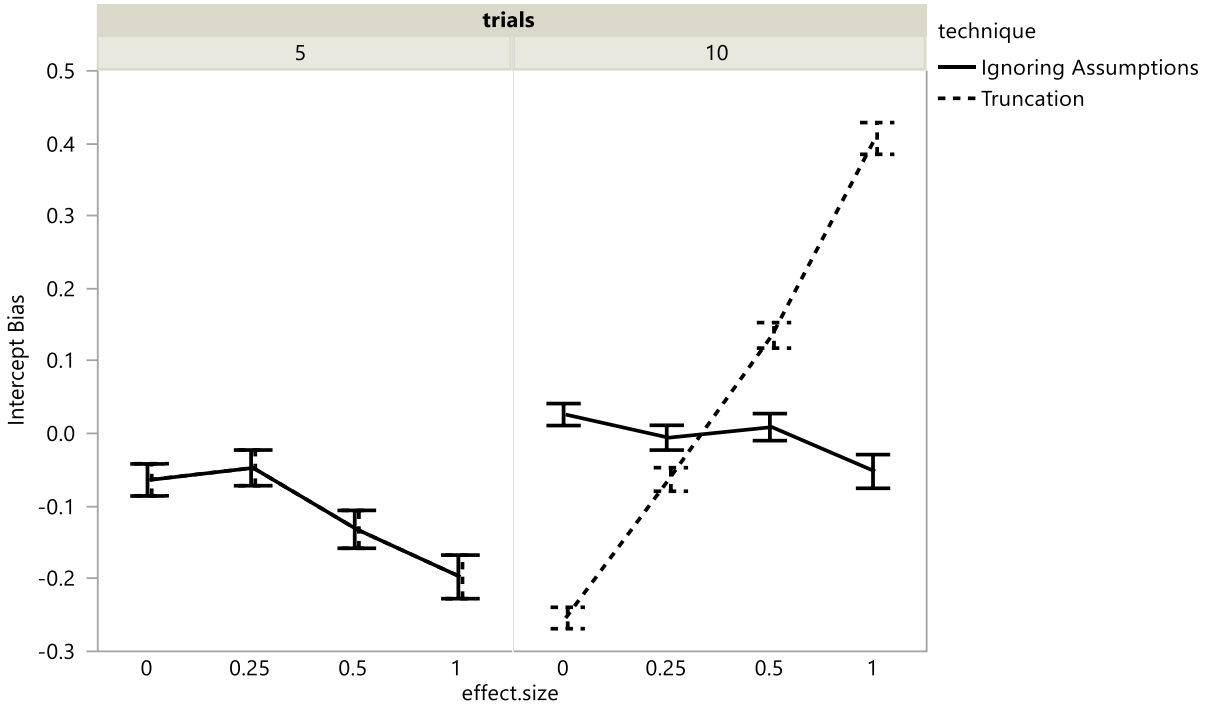

Figure 9 Bias in the intercept parameter estimate by effect size and number of trials per subject. Error in the intercept parameter estimates by effect size and number of trials per subject for continuous linear-scale predictors. Each error bar is constructed using 1 standard error from the mean.

Log scale.

When discrete log scaled predictors are used, applying the Gamma distribution (*M=-0.18, SD=0.564*), *t=-18.83, p<0.001,* and transforming the data (*M=-0.01, SD=0.056*), *t=-19.87, p<0.001,* caused underestimation. When continuous log scaled predictors are used, applying the Gamma distribution (*M=0.14, SD=1.14*) caused overestimation of the intercept parameter, *t=19.11, p<0.001,* while transforming the data (*M=-0.02, SD=0.12*) causes underestimation *t=-20.56, p<0.001*.

When measuring estimation bias in the intercept parameter estimates, aggregating and ignoring assumptions produced unbiased estimates when discrete predictors are used. However, truncating the data using discrete or continuous predictors causes significant underestimation of the intercept.

Difference parameter bias.

Linear scale.

When linear-scale discrete predictors were used, technique was not a good predictor of estimation bias in the difference parameter, but it did significantly interact with the number of trials per subject (Table 16). At 5 trials per subject, there were no differences between techniques, $F(2,2)=0.07$, $p=0.9320$. At 10 trials per subject, there were differences between techniques, $F(2,2)=8.84$, $p<0.001$ such that ignoring assumptions and aggregation were significantly different from truncation (Fig 10). Ignoring assumptions and aggregation did not significantly bias the difference estimate ($M=-0.004$, $SD=0.576$) but truncation caused overestimation of the difference parameter ($M=0.024$, $SD=0.597$) which corresponds to a smaller difference between the conditions. In the analysis, SAS computes the difference between groups as condition 1 – condition 2. Since the second condition is simulated to be larger than the first, the estimated difference will be negative. In this case, overestimating the difference parameter will actually result in a less negative value (i.e. a smaller difference between the means).

Table 16

Effects model for the difference parameter estimate

	Discrete Linear			Continuous Linear		
Source	DF	F	p-value	DF	F	p-value
Technique	2	2.71	0.0666	1	201.42	<.0001
Trials	1	15.74	<.0001	1	1936.00	<.0001
technique*trials	2	3.45	0.0317	1	201.42	<.0001
sample.size	2	0.66	0.5167	2	0.79	0.4557
technique*sample.size	4	0.01	0.9998	2	0.01	0.9908
trials*sample.size	2	1.72	0.1789	2	1.40	0.2473
technique*trials*sample.size	4	0.02	0.9995	2	0.01	0.9908
effect.size	3	0.31	0.8156	3	39.01	<.0001
technique*effect.size	6	0.62	0.7127	3	48.61	<.0001
trials*effect.size	3	2.75	0.0414	3	328.67	<.0001
technique*trials*effect.size	6	0.59	0.742	3	48.61	<.0001
sample.size*effect.size	6	1.69	0.1194	6	0.50	0.8052
technique*sample.size*effect.size	12	0.01	1	6	0.01	1
trials*sample.size*effect.size	6	4.17	0.0003	6	1.87	0.0824
technique*trials*sample.size*effect.size	12	0.03	1	6	0.01	1

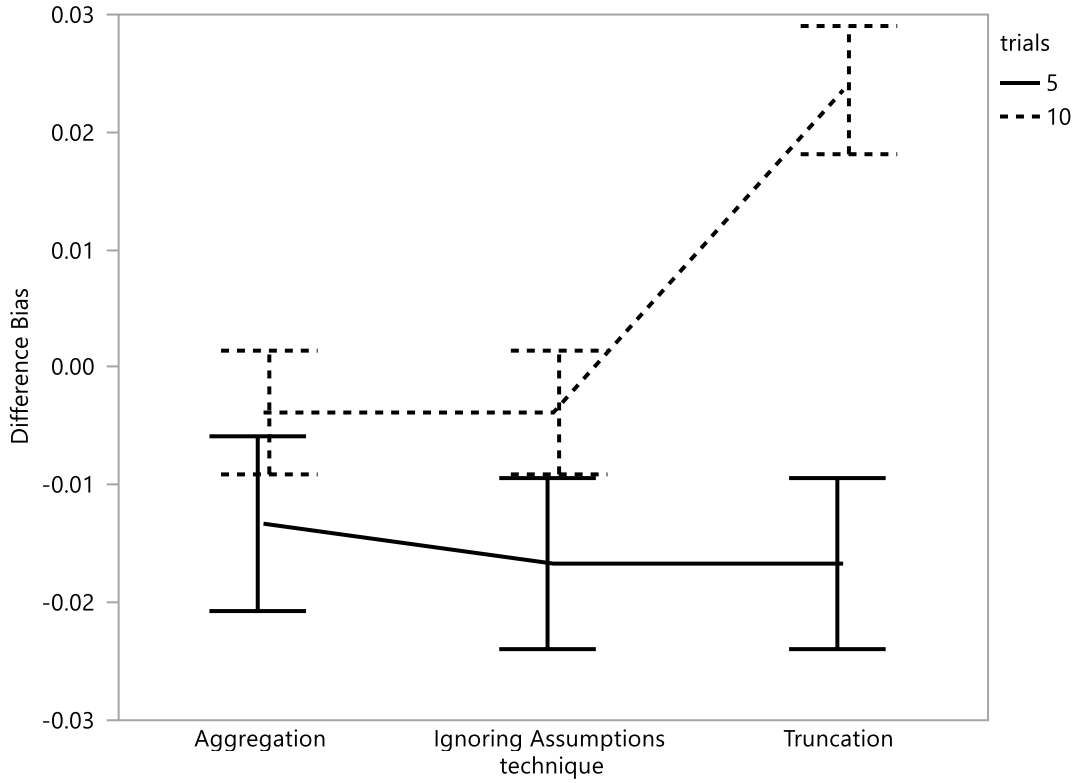

Figure 10 Bias in the difference parameter estimate by technique and number of trials per subject. Error in the difference parameter estimates by technique and number of trials per subject for discrete linear-scale predictors. Each error bar is constructed using 1 standard error.

For linear-scale continuous predictors, there was a significant difference among the techniques (Table 16). Ignoring distributional assumptions (*M=0.049, SD=0.36*) overestimates the slope parameter while truncation (*M=0.003, SD=0.37*) is virtually unbiased. Truncation maintained little to no estimation bias as effect size increases, but ignoring distributional assumptions led to increasing overestimation of the slope. When truncation did not occur (5 trials), both continuous linear-scale techniques overestimated the slope parameter estimate, but when truncation did occur (10 trials), ignoring assumptions was an unbiased method while truncation underestimated the slope estimates (Fig 11). These effects are both present in the

significant three-way interaction between technique, number of trials per subject, and effect size (Table 16).

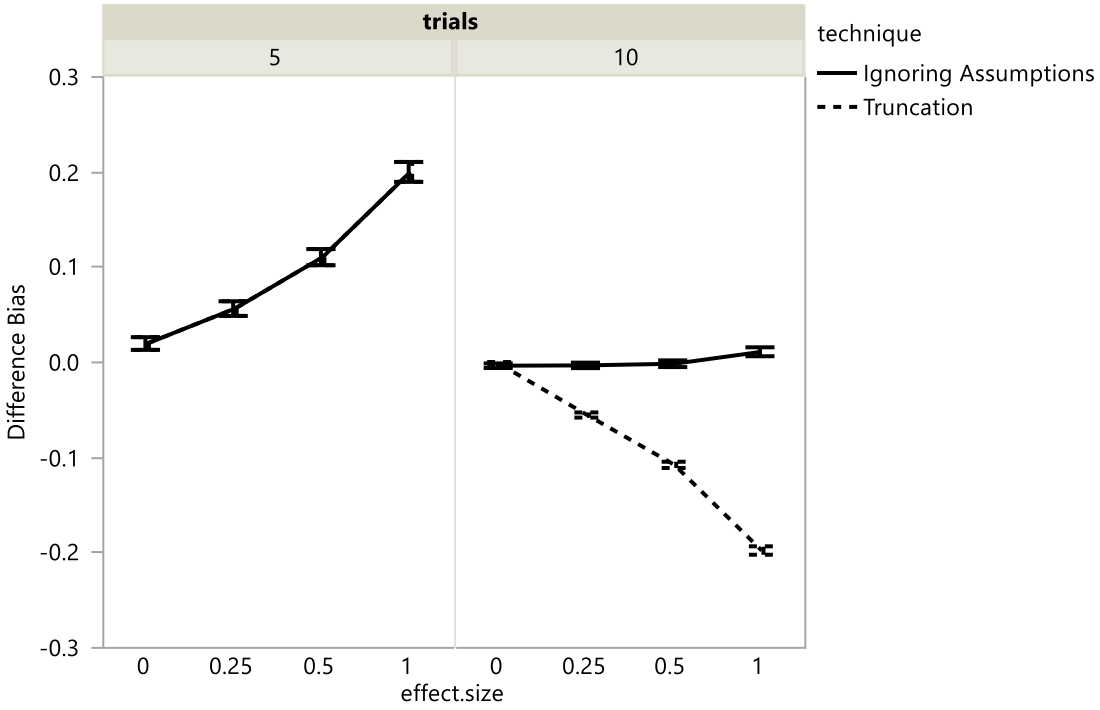

Figure 11 Bias in the difference parameter estimate by effect size and number of trials per subject. Error in the difference parameter estimate for continuous linear-scale predictors by effect size and number of trials per subject. Each error bar is constructed using 1 standard error from the mean.

Log scale.

When log scaled discrete predictors were used, applying the Gamma distribution (*M=0.002, SD=0.697*) did not cause significant estimation bias, but transforming the data (*M=-0.033, SD=0.566*) slightly overestimated the difference, *t=-9.41, p<0.0001*. In SAS, the difference is estimated by taking condition 1 – condition 2. In this case, a more negative difference value (i.e. underestimation of the difference parameter) actually corresponds to an increase in the difference between conditions. When log-scale continuous predictors were used, applying the Gamma distribution (*M=0.007, SD=0.029*) and transforming the data (*M=0.007,*

SD=0.034) both significantly overestimated the slope parameter, *t=38.34, p<0.0001* and *t=32.92*, respectively.

Overall, there appears to be some masking of the truncation effects. When truncation did not occur (at 5 trials), the analysis overestimated the difference parameter but equally underestimated it when truncation did occur (at 10 trials). This caused the overall effect of truncation to appear unbiased.

Intercept parameter accuracy.

Linear scale.

When discrete predictors were used, the analysis identified a significant difference between the techniques. When measuring the error in the intercept parameter for continuous linear-scale predictors, there was not a significant difference between techniques (Table 17).

Table 17

Effects model for the intercept parameter accuracy

	Discrete Linear			Continuous Linear		
Source	DF	F	p-value	DF	F	p-value
Technique	2	152.03	<.0001	1	0.12	0.7338
Trials	1	1010.44	<.0001	1	1960.87	<.0001
technique*trials	2	150.23	<.0001	1	0.12	0.7338
sample.size	2	3644.20	<.0001	2	2282.12	<.0001
technique*sample.size	4	5.18	0.0004	2	1.34	0.2614
trials*sample.size	2	63.26	<.0001	2	12.23	<.0001
technique*trials*sample.size	4	5.57	0.0002	2	1.34	0.2614
effect.size	3	273.09	<.0001	3	369.62	<.0001
technique*effect.size	6	0.07	0.9986	3	1.19	0.3117
trials*effect.size	3	6.86	0.0001	3	5.24	0.0013
technique*trials*effect.size	6	0.08	0.9978	3	1.19	0.3117
sample.size*effect.size	6	5.65	<.0001	6	6.96	<.0001
technique*sample.size*effect.size	12	0.14	0.9997	6	0.47	0.8286
trials*sample.size*effect.size	6	2.18	0.0419	6	2.26	0.0348
technique*trials*sample.size*effect.size	12	0.12	0.9999	6	0.47	0.8286

Truncating the data was significantly less accurate than aggregating the data or ignoring distributional assumptions when discrete predictors were used (Table 18). As effect size

increased, all techniques became less accurate, but truncating the data was far less accurate than the other two techniques. At 5 trials, when truncation did not occur, all techniques performed the same. With 10 trials, where truncation did occur, aggregating the data and ignoring distributional assumptions greatly improved in accuracy while truncating the data maintained inaccurate estimates of the intercept parameter (Fig 12). These effects were present in the significant three-way interaction between technique, sample size, and number of trials per subject (Table 17).

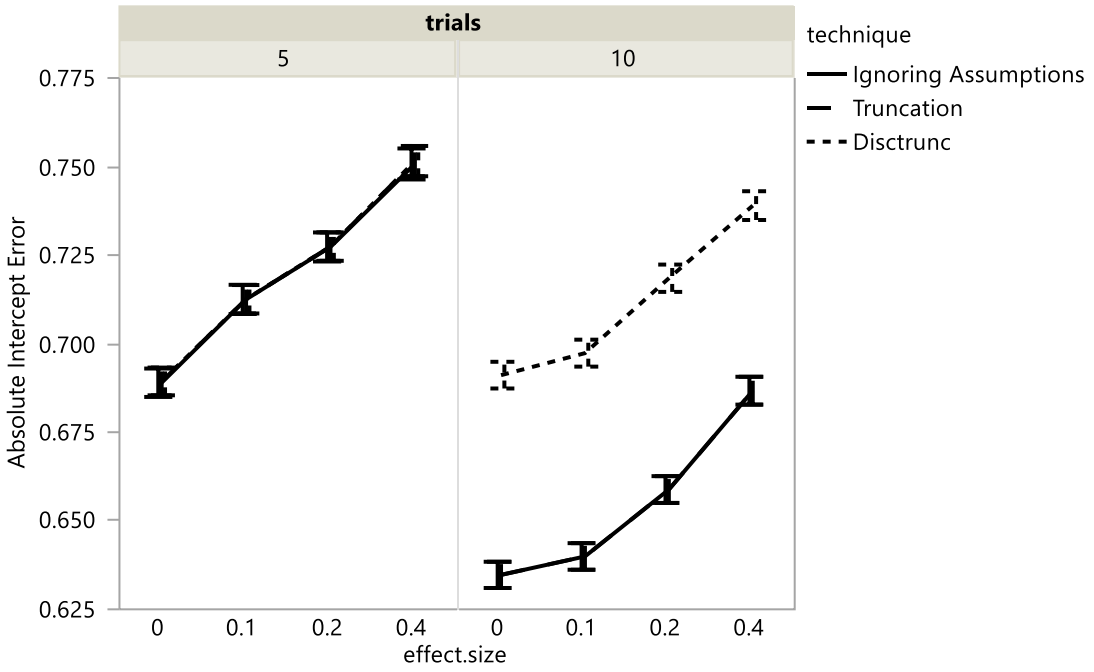

Figure 12 Absolute error in the intercept parameter estimate by effect size and number of trials per subject. Accuracy in the intercept parameter estimates by effect size and number of trials per subject for discrete linear-scale predictors. Each error bar is constructed using 1 standard error from the mean.

Table 18

Post hoc means comparisons for technique for discrete predictors. Techniques with the same letter are not significantly different from each other.

Technique			Least Sq Mean
Aggregation	A		0.687
Ignoring Assumptions	A		0.688
Truncation		B	0.716

Log scale.

When log scaled continuous predictors were used, both applying the Gamma distribution (*M=0.88, SD=0.279*) and transforming the data (*M=0.414, SD=0.135*) had significant error when estimating the intercept parameter, *t=343.46, p<0.001* and *t=332.8, p<0.0001*, respectively. Similar patterns emerged when log-scale discrete predictors were used. Applying the Gamma distribution (*M=0.706, SD=0.222*) and transforming the data (*M=0.32, SD=0.103*) were significantly inaccurate, *t=350.86, p<0.0001* and *t=342.97, p<0.0001*, respectively.

For the intercept parameter, truncating the data and ignoring distributional assumptions were not different in their accuracy of the estimates when continuous predictors were used but did differ when discrete predictors were used. For the latter, truncation was significantly less accurate than other methods.

Difference parameter accuracy.

Linear scale.

When linear-scale discrete predictors were used, technique was not a significant predictor of the difference parameter estimate. When measuring the error in the difference parameter for linear-scale continuous predictors, there was a significant difference between techniques (Table 19).

Table 19

Effects model for the difference parameter accuracy

	Discrete Linear			Continuous Linear		
Source	DF	F	p-value	DF	F	p-value
Technique	2	2.93	0.0534	1	32.22	<.0001
Trials	1	2867.92	<.0001	1	11387.87	<.0001
technique*trials	2	2.71	0.0662	1	32.22	<.0001
sample.size	2	4027.62	<.0001	2	1811.37	<.0001
technique*sample.size	4	0.07	0.9915	2	8.87	0.0001
trials*sample.size	2	34.34	<.0001	2	110.28	<.0001
technique*trials*sample.size	4	0.06	0.9938	2	8.87	0.0001
effect.size	3	185.66	<.0001	3	1195.72	<.0001
technique*effect.size	6	0.02	1	3	18.54	<.0001
trials*effect.size	3	6.28	0.0003	3	3.38	0.0173
technique*trials*effect.size	6	0.02	1	3	18.54	<.0001
sample.size*effect.size	6	14.37	<.0001	6	3.98	0.0005
technique*sample.size*effect.size	12	0.06	1	6	2.54	0.0185
trials*sample.size*effect.size	6	5.17	<.0001	6	2.92	0.0075
technique*trials*sample.size*effect.size	12	0.07	1	6	2.54	0.0185

Both techniques had significant error when estimating the difference, but truncating the data (*M=0.615, SD=0.19*) was less accurate than ignoring distributional assumptions (*M=0.606, SD=0.192*). For small effect sizes, 5 trials (no truncation) and small sample sizes, all techniques had similar accuracy in estimating the difference parameter. But as sample and effect sizes increase and with 10 trials, truncating the data had significantly less accurate estimates of the difference parameter than ignoring distributional assumptions (Fig 13).

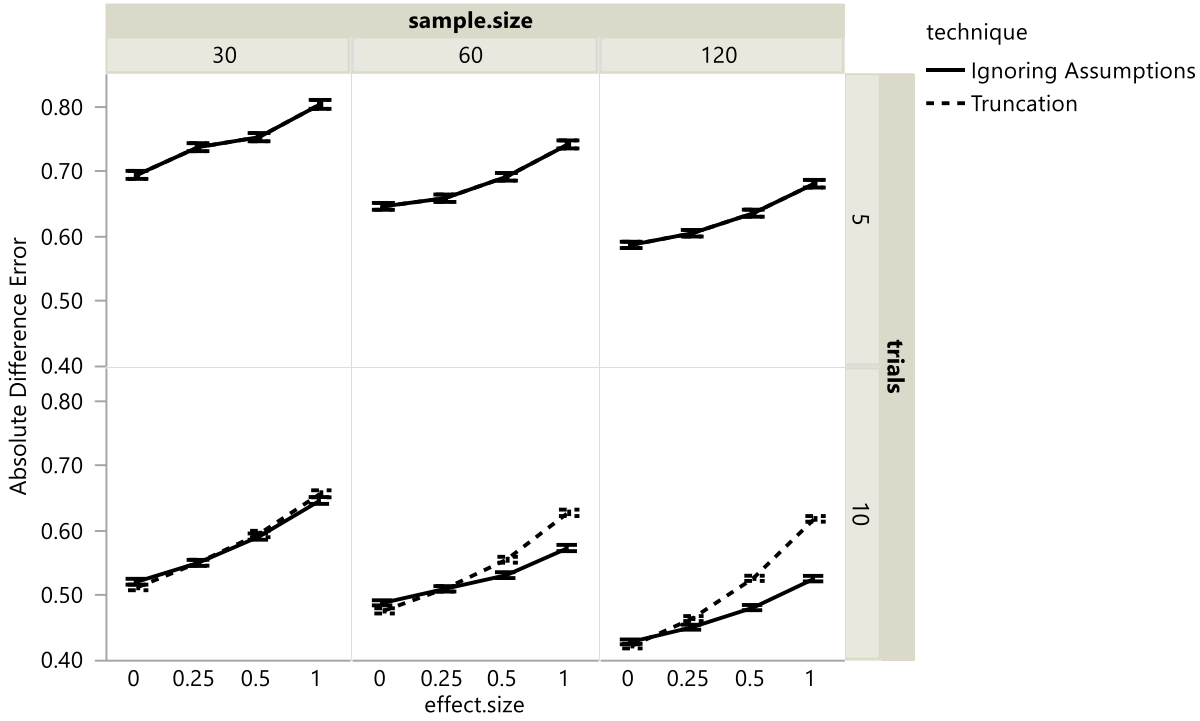

Figure 13 Absolute error in the difference parameter estimates by effect size, number of trials per subject, and sample size for continuous linear-scale predictors. Each error bar is constructed using 1 standard error from the mean.

Log scale.

When log scaled continuous predictors were used, applying the Gamma distribution (*M=0.296, SD=0.096*) and transforming the data (*M=0.308, SD=0.099*) were significantly inaccurate, *t=284.14, p<0.0001* and *t=286.56, p<0.0001*, respectively. Both applying the Gamma distribution (*M=0.765, SD=0.215*) and transforming the data (*M=0.722, SD=0.203*) produced significant error when estimating the difference, *t=339.92, p<0.0001* and *t=340.01, p<0.0001*, respectively, when log-scale discrete predictors were used.

Discrete linear-scale predictors estimated the difference parameter with similar accuracy. When estimating the difference parameter using continuous predictors, all techniques were significantly inaccurate, but truncation was significantly less accurate than ignoring assumptions.

Chapter 4- Discussion

Overall, this research aimed to identify the effects that several common techniques have on analyzing reaction time data. The results of this research partially supported the presented hypotheses. Applying the Gamma distribution and log-transforming the data were expected to perform the best, having the most power to identify effects, a criterion that maintains a 5% Type 1 error rate, and unbiased and accurate parameter estimates. This hypothesis was not supported, as both techniques produced biased intercept estimates and the log transformation had low discriminability (i.e. poor ability to identify effects). Truncating the data was expected to perform poorly. Prior research in the field (Ulrich & Miller, 1994; Baayen & Milin, 2010) and the results of preliminary research indicated that truncating the data would have poor discriminability and underestimates condition means (Tables 2 & 3). This hypothesis was supported. It was expected that ignoring distributional assumptions and aggregating the data would have poor discriminability and low accuracy for parameter estimation (Lorch & Myers, 1990), and this hypothesis was not supported. Each technique is discussed in detail.

There were several consistent results, regardless of technique applied. In general, more observations resulted in higher levels of discriminability, more neutral criterion values, and increased parameter accuracy. Larger effects sizes generally resulted in more biased parameter estimates and less parameter accuracy.

Impact of Ignoring Distributional Assumptions

Ignoring distributional assumptions can sound like a technique to be avoided but, overall, it was not that different from the other techniques when using discrete predictors. When continuous predictors were used, this technique had high discriminability and a neutral criterion (Table 7). This is not surprising since the simulated effect sizes were chosen to show a clear effect. The medium and large effect sizes were so distinct that when looking at all effect sizes

together, ignoring distributional assumptions did not seem to decrease the discriminability. As preliminary results suggested, ignoring distributional assumptions did conservatively bias the criterion value (Table 7). This conservative criterion will cause the technique to miss some significant effects in the data.

For both types of predictors, both the intercept and difference parameters were unbiased. To help with visualizing this estimation bias, the actual and predicted values are superimposed on violin plots created from 100,000 simulated observations (Figs 14 and 15). Some estimation bias in the slope estimate emerges when a continuous predictor is used (Fig 15), but not significantly so. When data are skewed, there is more variability around the longer RTs and the skew begins to flatten (Fig 16). These longer tails will pull the estimated means up if the skew is ignored, creating a steeper slope estimate.

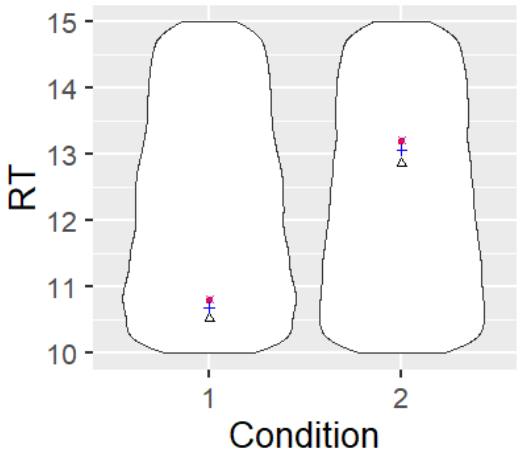

Figure 14 Results from simulating 100,000 observations for a discrete predictor, zoomed in to focus on the different effects. Circular dots represent the theoretical means. The "x" symbol represents the predicted means from the earlier analyses when distributional assumptions are ignored or RTs were aggregated. Parameter estimates are identical for ignoring assumptions and aggregation. Triangles represent the predicted means when data are truncated. Crosses represent the predicted means when a Gamma model is used.

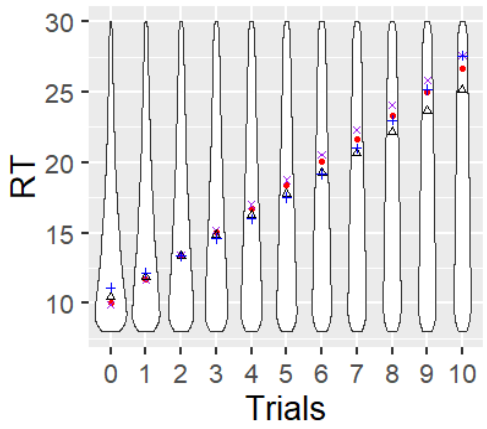

Figure 15 Results from simulating 100,000 observations for a continuous predictor zoomed in to focus on the different effects. Circular dots represent the theoretical means. The "x" symbol represents the predicted means when distributional assumptions are ignored. Triangles represent the predicted means when data are truncated. Crosses represent the predicted means when a Gamma model is used.

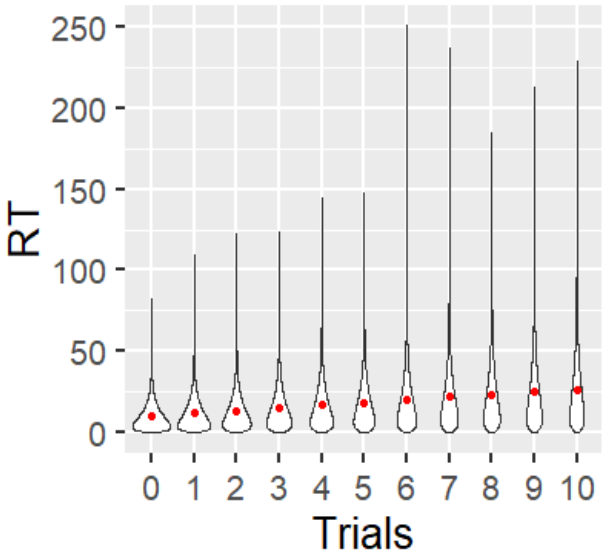

Figure 16 Results from simulating 100,000 observations for a continuous predictor. Dots represent the theoretical means at each trial. As trials increase, the distributions become flatter and more spread out.

However, ignoring assumptions starts to break down for extreme conditions (i.e., few observations and large effect sizes), particularly for continuous predictors (Figs 9-12). When there are a small number of observations, there will be more uncertainty and less precision. Ignoring distributional assumptions causes significant underestimation of the intercept parameter for both types of predictors when there are only 5 trials per subject. This means the tails of the originating distribution would be longer, increasing the mean response and causing overestimation of the difference when the skew is ignored. This would tend to drive the intercept estimate down. For continuous predictors, in particular, larger effect sizes and a smaller numbers of observations led to significantly inaccurate estimates of both the intercept and the slope. When the effect was large, ignoring assumptions will be more impactful and noticeable than when the effect is smaller. Fitting a linear model to skewed data will make it difficult to assess the parameters, particularly the intercept, because the variability around each point is misattributed to be equally distributed. With skewed data, there is more variability around larger values than smaller ones.

Ignoring distributional assumptions might be acceptable if an experiment requires high discriminability and only uses discrete predictors. However, if continuous predictors are used, the benefit of the high discriminability will be greatly offset by the inaccurate parameter estimates. If one is simply interested in unbiased estimates, ignoring the distributional assumptions could be used, but only when there is a large number of observations (i.e. n=120). This technique should be avoided if the research involves a small number of trials per subject or a small number of subjects. It should also be avoided if the research is interested in the accuracy of parameter estimates and uses continuous predictors.

Impact of Aggregation

Aggregating the data produced similar results to ignoring the distributional assumptions, a testament to the central limit theorem. All sample sizes were greater than 30, so there were a sufficient number of observations so that the means were normally distributed. This produced the same results as simply assuming the normal distribution to begin with.

Aggregating the data is a common technique used by researchers. Because general trends are typically of interest, collapsing data across some variables may seem effective. However, when data are skewed, the mean is not the best measure of central tendency. Analyzing only the means ignores the skew in the data and can affect the results in different ways (Lorch & Myers, 1990). In this research, aggregating the data did not have an especially detrimental effect on discriminability or criterion (Table 7). When compared to the other discrete predictors, aggregating the data was among the techniques with the highest discriminability. Again, this result is likely due to the simulated effect sizes. With the large effects, the amount of overlap between distributions may be similar whether they come from all the observations or the average of each participant. The slight conservative bias in the criterion value is most likely because of the upward push of the means when aggregated. Although the mean estimates were not significantly different from the theoretical values, the effect still shows up as a slightly conservative criterion value. This conservative criterion will cause the technique to miss some significant effects in the data

Among the linear-scale techniques, aggregating the data produced the least amount of estimation bias and the most accurate intercept parameter estimates (Table 20) and was nearly identical to ignoring the assumptions. This effect was consistent across all sample sizes, number of trials per subject and effect sizes.

Table 20

Means for the intercept bias and accuracy for discrete linear-scale predictors

	Ignoring Assumptions	Aggregation	Truncation
Estimation Bias	0.007	0.005	-0.138
Absolute Error	0.688	0.687	0.716

The performance of aggregation is likely due to the amount of simulated skew generated for a discrete predictor. The effect that the skew has on the means starts to become apparent when using a continuous predictor. Since aggregation across trials can only occur when there are replications at each level of a predictor, the differential effect of using only means is not observable with the continuous predictor used in this research where there were no within-subject replications. The skew is not very different between the two conditions (Fig 17). In this case, the amount of estimation bias in the means would be about equal in both conditions, which would result in unbiased difference parameter estimates.

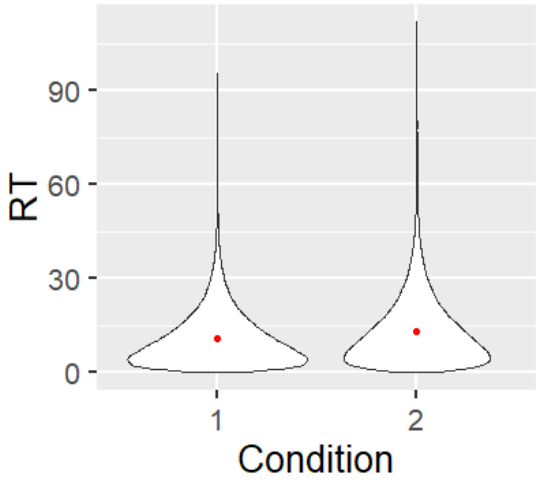

Figure 17 Results from simulating 100,000 observations for a discrete predictor. Dots represent theoretical means.

Impact of Transformation

Transforming the data is a good way to produce normality in a skewed dataset. A common transformation for RT data is applying a logarithmic function to the raw data and analyzing them using linear-scale techniques assuming a normal distribution. This method works well when the log transformation truly produces normality, but it does not always do so. When I examined the effect of the transformation, it created a resemblance of normality. However, the data become slightly skewed in the opposite direction (Fig 18) and the residuals were still not normally distributed (Fig 19).

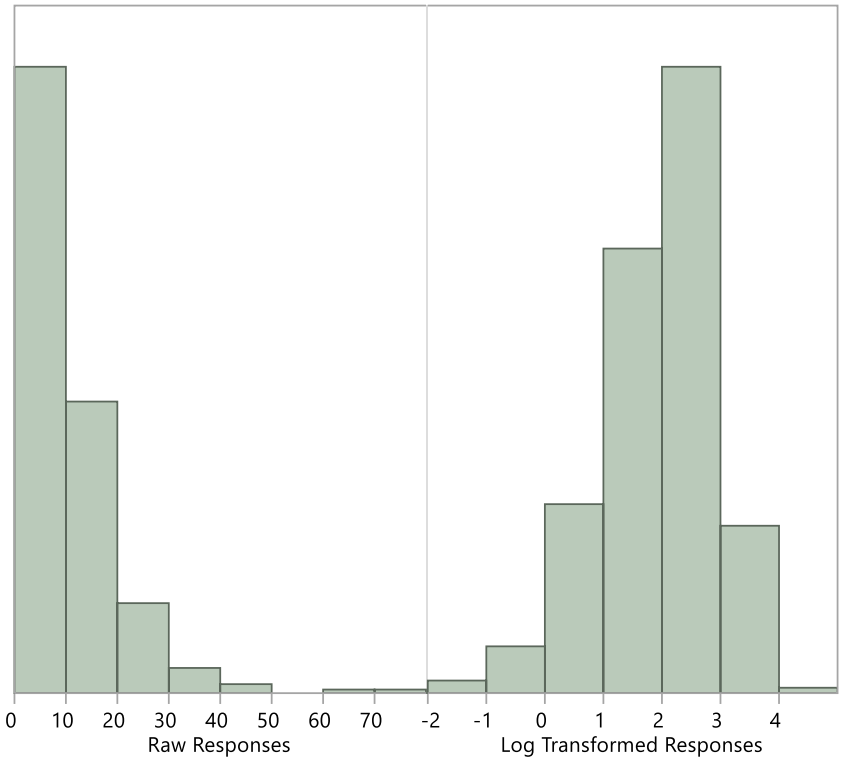

Figure 18 Histograms of raw (left panel) and transformed (right panel) data.

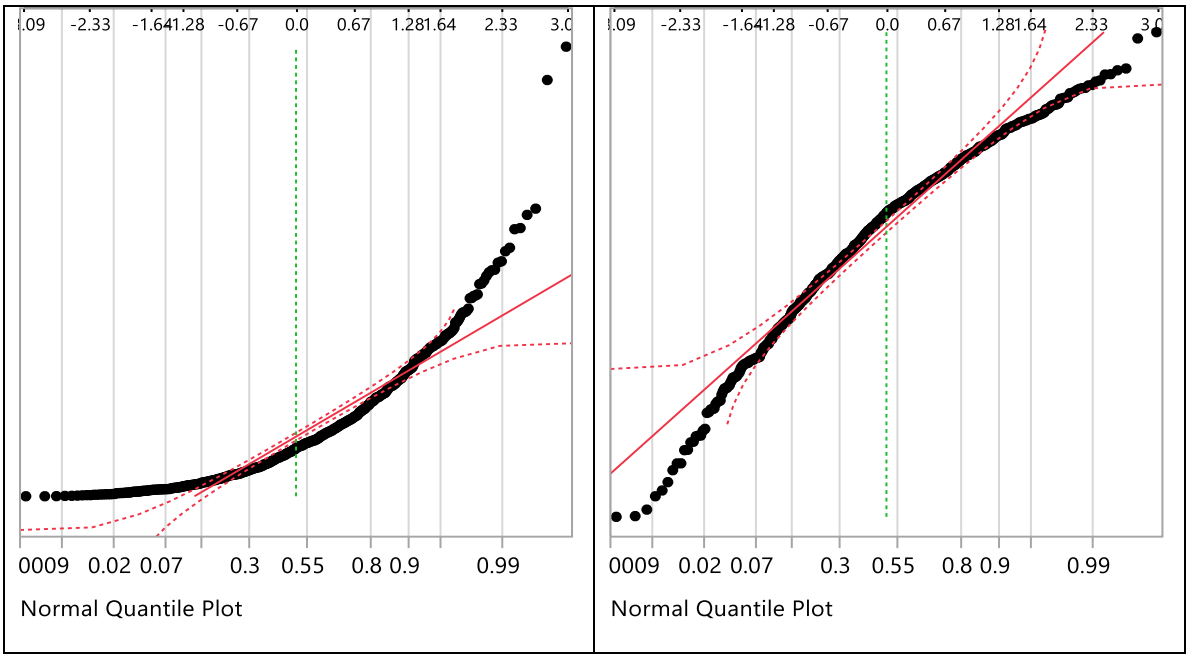

Figure 19 Q-Q plots of the residuals for raw (left panel) and log transformed (right panel) responses.

Transforming the data was the only technique that was significantly different from the rest in terms of discriminability. It was significantly worse than other techniques at discriminating a significant effect from a non-significant one. As Figure 18 shows, transforming the data spread out the distribution quite a bit. In fact, it caused the distribution to become slightly left-skewed. This over-compensation increased the overlap between the distributions and greatly minimized discriminability. This overlap will cause more errors, both Type 1 and Type 2.

Transforming the data did not create bias in either parameter estimate when continuous predictors were used with 10 trials (Figure 20). This effect was consistent across all effect sizes and sample sizes. Accuracy was always better when there were more observations. Intercept parameter estimates were more accurate when discrete predictors were used, whereas the difference parameter estimates were more accurate with continuous predictors.

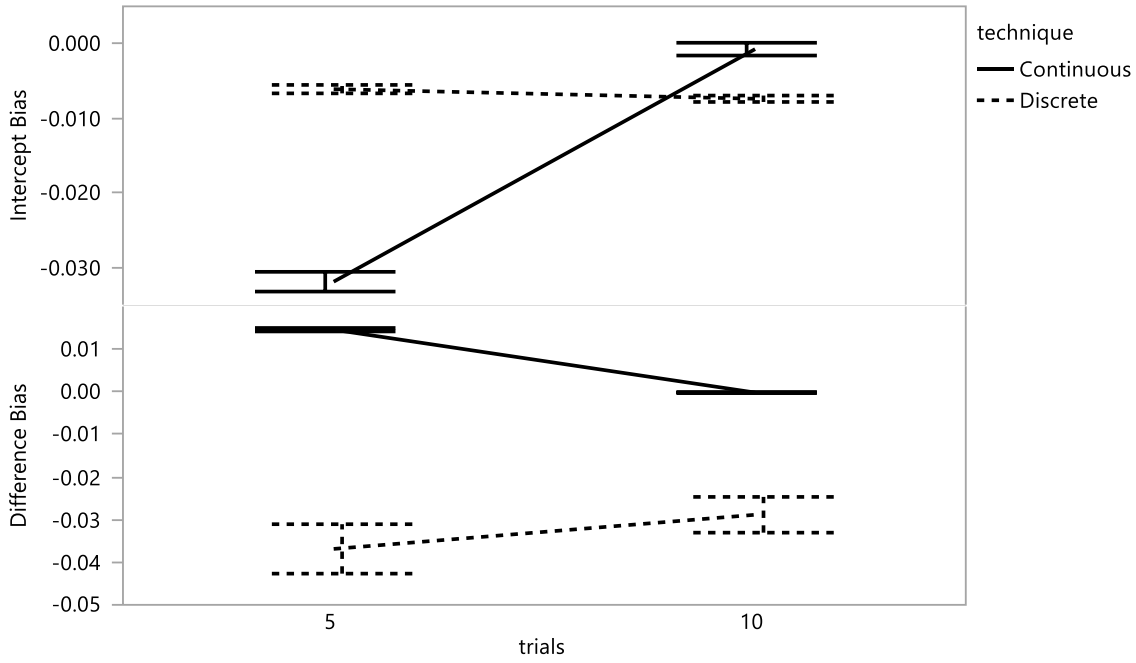

Figure 20 Bias in the intercept and difference parameter estimates for continuous and discrete predictors when data are transformed.

When discrete predictors were used, skewing the data in the opposite direction caused the intercept parameter estimates to become underestimated, while overestimating the difference between the means (Fig 21). This is the case for both 5 and 10 trials per subject.

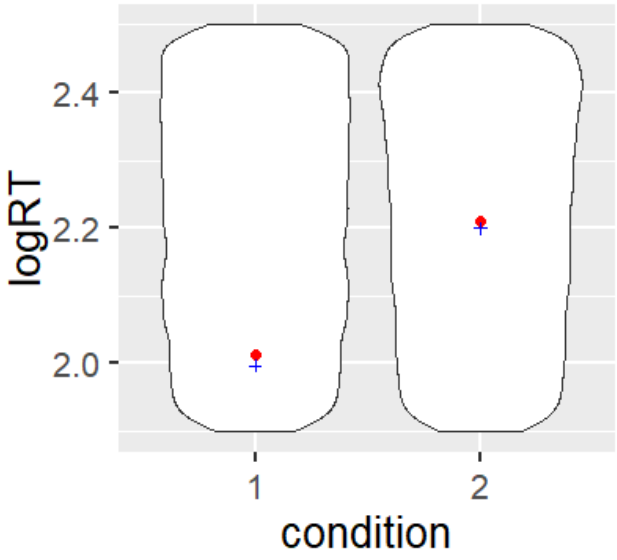

Figure 21 Results from simulating 100,000 log transformed observations for a discrete predictor, zoomed in to focus on the different effects. Circular dots represent the theoretical means. The cross symbol represents the predicted means when a normal distribution is fit to the log transformed data.

When a continuous predictor is used with 10 trials per subject, the intercept and difference estimates were generally unbiased. The original means are curved because log transforming the data skews the distributions in the opposite direction. This causes the means to increase at a decreasing rate as the number of trials increases (Fig 22). Although the fitted line represents the optimal fit to this curve (i.e., there was a non-significant difference between the estimated and true intercepts), this represents a model misspecification in which a curved relationship is fit with a straight line. The consequences of this misspecification are more apparent for larger effect sizes. As effect size increases, the curve will become steeper. Linearizing the relationship will cause an increasingly shallow regression line, leading to increasing underestimation of the slope. Estimation bias is introduced for smaller numbers of trials per subject because there is more error variability around the estimates. Both the intercept

and slope are underestimated at 5 trials per subject (Fig 20). As the number of observations increases, the estimation bias diminishes.

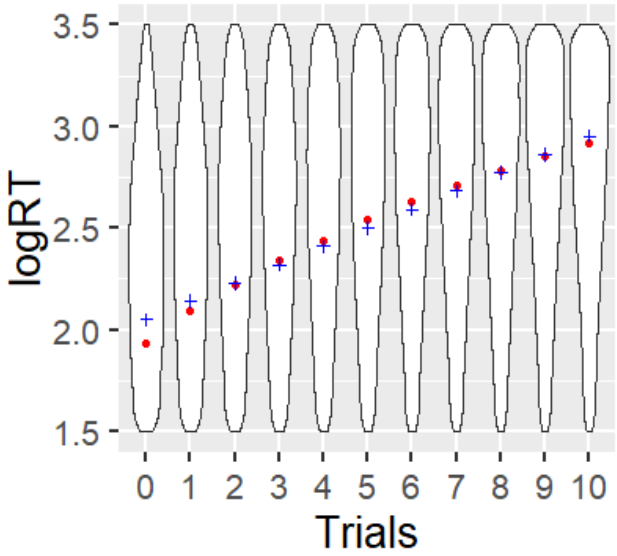

Figure 22 Results from simulating 100,000 log transformed observations for a continuous predictor, zoomed in to focus on the different effects. Circular dots represent the theoretical means. The cross symbol represents the predicted means when a normal distribution is fit to the log transformed data.

Transforming the data can be useful when there are a sufficient number of trials per subject and continuous predictors are used. If there are few trials per subject, the impact of effect size decreases the efficacy of the transformation. Furthermore, if the research uses discrete predictors, the difference parameter estimates will be too inaccurate to justify transforming the data with a logarithmic function. One should also take caution interpreting the results of an analysis involving transformed data unless normality has been produced with certainty. For RT data, the log transformation may over-correct the skew and fail to establish normality.

Impact of Truncation

Truncating data is another very common analytic technique for analyzing skewed data. This technique seems to make sense on the surface. It is commonly assumed that observations far away from the average value, such as extremely long reaction times, represent some other process and should not be included with the rest of the data. These observations are then truncated so that a more normally distributed dataset remains. This method, however, has dire consequences, particularly when those extreme values could be valid observations (Baayen & Millin, 2010; Ratcliff, 1993; Ulrich & Miller, 1994).

This technique had reasonable discriminability and criterion (Table 7). Some of these effects were masked when looking at overall technique effects. Including 5 trials per subject (when truncation did not occur) often minimized the effects of the truncation. This is why comparative results were looked at within the number of trials per subject. When we only look at 10 trials per subject and focus the analysis on when truncation occurred we find that truncating the data according to the mean resulted in significantly lower discriminability and significantly liberal criterion. This will cause the technique to make more Type 1 errors and identify effects that are not truly there.

When truncation occurs with a discrete predictor, the upper tails within each condition are truncated. This will result in differential truncation because the skew is different for each condition. There is more skew in the lower values than the higher values, so the truncation will be more in the first condition than the second. This will drive the intercept downward. Since truncation still occurs in both conditions, the difference will be underestimated (Fig 14). The effects of truncation are more noticeable when a continuous predictor is used. In this case, truncation occurs within subject so there is one cut-off value and everything about that value is omitted. This results in more truncation of the upper tail; more truncation occurs as the number

of trials increases. This will cause the slope to flatten and driving the intercept upward. The result is overestimation of the intercept and underestimation of the slope (Fig 15) (Ulrich and Miller, 1994).

Truncation should be avoided. This method greatly inflates the Type 1 error rate and produces far more biased and less accurate parameter estimates than other methods. Parameter estimates were also consistently biased and inaccurate when the effect size was large and when the number of subjects was small. Its gross underestimation of the difference parameter makes this a poor technique to use, as this is often the parameter of interest for researchers.

Impact of Applying a Gamma Distribution

Modeling the data with the distribution from which it was generated should be the best method, although determining the generating distribution from actual data can be challenging. The aim of this research was to determine the conditions in which tackling this challenge would be beneficial. In this case, the appropriate model is the Gamma distribution because the data were simulated from a Gamma distribution. However, it is also important to identify the true link function to determine relationships in the data. In the present research, the data were generated such that the means were linear, so fitting a distribution using the log link function produced a model misspecification. In reality, one would need to determine the best distribution and link function to model the data.

Although applying the Gamma distribution to the data was not statistically different from other methods in signal detection, except for transformation, it did have the highest discriminability and a neutral criterion value when a discrete predictor was used (Table 7). This technique provided the strongest distribution model for accurately detecting significant results. The effects of applying the Gamma distribution on parameter recovery depends on the type of

predictor. When a discrete predictor is used, applying the Gamma distribution underestimates the intercept (Fig 14) while it overestimates it when a continuous predictor is used (Fig 15). However, for a discrete predictor, applying the Gamma distribution does not bias the difference estimate. Appropriately identifying the skew in the data allows the Gamma model to accurately describe the effect. The misestimation of the intercept is likely due to the linear pattern of the means in the simulated data. Applying the Gamma model assumes a curvilinear pattern of means, so the effects on the intercept will be different depending on the type of predictor. When a discrete predictor is used, it is not clear why group means were underestimated because having the incorrect link function should have little effect under these conditions. When a continuous predictor is used, the intercept will be overestimated because the analysis is fitting a curved pattern of means such that the means increase at an increasing rate whereas the actual means are increasing at a constant rate.

There was some variability in the intercept bias that can cause some bias in the estimates, but the difference estimates were unbiased for both trial sizes and all sample sizes. The same is true for small and medium effect sizes. The difference parameter estimate was not impacted by effect size when there were 10 trials. When discrete predictors were used, difference estimates were consistently unbiased across the number of trials per subject and effect size. The intercept parameter estimates were more accurate when discrete predictors were used and the slope parameter estimates were much more accurate when continuous predictors were used.

Effect size had a large impact on estimation bias for both parameter estimates. When continuous predictors were used, this effect was greatest for the intercept estimates when there were 10 trials and greatest for the difference estimates when there were only 5 trials. This variability is likely caused by the estimation technique. This research used the generalized linear

model using PROC GLIMMIX specifying a Gamma distribution with a log link function in SAS® 9.4. Using a different link function, such as the inverse, or another procedure, such as PROC GENMOD could minimize this variability. Furthermore, the shape parameter was fixed in the generation of the data, but not fixed in the analyses. This additional parameter estimation will likely increase the estimation biases.

Surprisingly, applying the Gamma distribution is not effective in estimating the intercept parameter. Oftentimes researchers are most interested in estimating the difference or the slope. In this case, applying the Gamma distribution appears to be an unbiased and relatively accurate option, regardless of predictor type.

Impact of the number of observations.

Increasing the number of observations reduces the parameter variance. Reducing the error variance will allow a test to more accurately detect significant differences, thereby improving discriminability and bias. As the number of observations increase, the margin of error decreases, making our estimates more accurate as well. More observations will also help determine the underlying distribution, if a generalized linear model is to be used.

However, there is one condition, truncation, where increasing the number of observations does not necessarily improve discriminability, estimation bias and accuracy. When the number of observations increased by increasing the number of trials per subject, the analysis deteriorated. This is because truncation occurred only at 10 trials per subject, so the effects of truncation were not seen at 5 trials. Truncating the data caused the parameter recovery to be no better than only having a few trials per subject. The other linear-scale techniques, ignoring assumptions and aggregation, improved in accuracy when the number of trials per subject increased. When truncation occurred, the accuracy was the same as ignoring assumptions and aggregation at only

5 trials per subject. None of the methods were differentially impacted by a change in the sample size.

Impact of effect size.

Increasing the effect size has a positive impact on discriminability and criterion. When the effect is small, it is more difficult for a test to discriminate between the noise and signal distributions, but as the effect size increases, this distinction becomes more apparent. The amount of overlap between the two distributions is minimized by increasing the effect size, thereby reducing any estimation bias. Increasing the effect size increases bias and decreases accuracy in the both parameter estimates for continuous predictors. Larger effect sizes have a stronger slope, so the effects of misestimating this relationship will be more apparent. Additionally, as the slope increases there will be more error variability. This will cause the estimates to be less accurate for larger effect sizes than for smaller ones.

Limitations

This research is limited in its ability to compare all techniques. Having techniques that use different scales do not allow direct comparison of parameter bias or accuracy. The research would be greatly improved by having a standardized measure of estimation bias and accuracy of the parameter estimates.

Another limitation is in the generation of the data. This data was generated such that the pattern of means was linear across the continuous predictor (Fig 23). In the field, this relationship is not usually linear. It tends to be curvilinear, following a more exponential relationship with differences becoming smaller as the data approach the RT floor (usually around 200 ms). In order to investigate the effects of analytical techniques on RT data commonly found in the field, a non-linear pattern of means would need to be generated. The linear pattern of

means caused the techniques that assumed a nonlinear relationship (i.e., the analysis based on log-transformed data and the Gamma regression with a log link function) to fare worse than they would with actual RT data due to model misspecification.

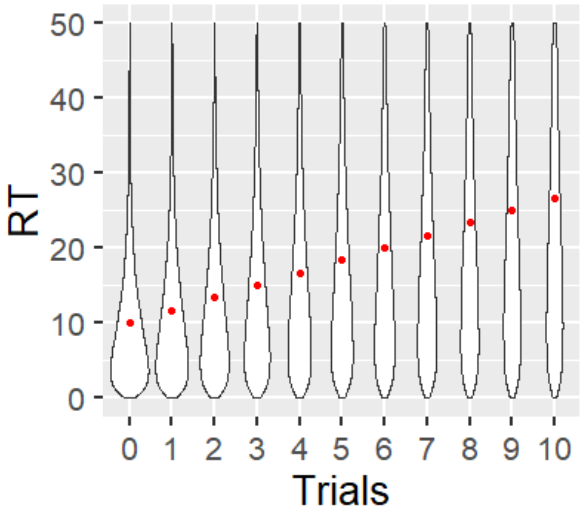

Figure 23 Results from simulating 100,000 log transformed observations for a continuous predictor. Dots represent theoretical mean values. Although the distributions are skewed and flattening as trials increases, the pattern of means is linear.

Specific Recommendations

1. Do not truncate. It reduces discriminability and produces a liberal criterion. This will cause the technique to discover non-existent significant results. When truncating RTs, only the longer observations are removed. This differential truncation causes distortion in the relationship between the predictor and the RTs and introduces bias in the parameter estimates.
2. Understand the relationship between the predictor and the average RTs. If the relationship is not obviously curvilinear, linear-scale models, such as aggregation or

ignoring assumptions, may perform just as well or better than applying a generalized linear model or transforming the data. Identifying the generating distribution may not be worth the effort when a simpler model could be used.

3. Consider the type of predictor being used. Discrete predictors are more robust because the relationship is much simpler. When using a continuous variable, it is important to properly identify any nonlinearities in the relationship between the variables.
4. When using a transformation, make sure it produces normality. Using an inappropriate transformation can decrease discriminability and produce an over conservative criterion. This would mean the transformation causes some significant results to be missed. Furthermore, the use of nonlinear transformations changes the assumed relationship between the predictors and the outcome (e.g., that it is exponential in the current study), and this assumed relationship should be appropriate for the data patterns observed.

Future Directions

The first step to improve this research is to find a direct comparison between the techniques for parameter estimation. This study would be more informative if it could include comparisons between the linear-scale and log-scale techniques to see which one truly is superior in estimating parameters. Furthermore, this research only studied one level of skew. These results may differ depending on different levels of skew. If the dataset were less skewed, the log-scale models may not be as advantageous over linear-scale models. In the future, it would be good to consider the effects of different analytical techniques on multiple levels of skew. This could help generalize this research to more than just RT data. There are also several other generalized linear models to be considered. Commonly used non-linear models are the ex-Gaussian and Weibull models (Van Zandt, 2000). Empirically, these models seem to fit reaction time data nicely. It would be a fine direction to study the effects of these models on signal

detection and parameter recovery. Perhaps investigating an exponential transformation with different shape parameters of the simulated data. Since the exponential distribution is in the Gamma family, one of the exponential transformations may be the correct one. It would also be beneficial to see the impacts of the number of observations and different effect sizes. Some other transformations to investigate could be the square root and the inverse transforms (Bartlett, 1947). These often produce normality and, given different levels of skew, may be more appropriate than the logarithmic transformation. One last direction to consider is censoring. This is a much better option than truncating (Dolan, van deer Maas & Molenaar, 2002). Instead of recording all observations and truncating those deemed outliers, the researcher could stop the trial at a specified duration to ensure unwanted trials are not included.

Overall, this research has shown the effects of different analytical techniques on reaction time data. The research should continue in order to find the most appropriate model for different conditions. The most appropriate model may not necessarily be the best fitting model. It may be the model that performs well enough. Sometimes the cost of finding the best fitting model may not be worth the benefit it provides. The field relies on accurate analyses which truly depend on many factors, including those investigated in this research. The suggestions listed provide a strong base to improve statistical analyses of RT data.

References

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3(2),* 12-28.

Bartlett, M. S. (1947). The use of transformations. *Biometrics, 3(1),* 39-52.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1).* 155-159.

Cutting, J.E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology, 44*. 3-19. DOI: 10.1006.jmps.1999.1274

DeCarlo, L.T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3(2),* 186-205.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59(4),* 447-456.

Dolan, C. V., Van der Maas, H. L., & Molenaar, P. C. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods, Instruments, & Computers, 34(3),* 304-323.

Erceg-Hurn, D.M., Mirosevich, V.M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63(7),* 591-601.

Jahn-Eimermacher, A., Lasarzik, I., & Raber, J. (2011). Statistical analysis of latency outcomes in behavioral experiments. *Behavioral Brain Research, 221(1),* 271-275.

Lawless, J. & Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis, 10,* 213-227.

Lo, S. & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6(1171)*

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(1),* 149.

Macmillan, N. & Kaplan, H. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98(1),* 185-199. DOI: http://dx.doi.org/10.1037/0033-2909.98.1.185

Macmillan, N. & Creelman, C. (1991). "Detection Theory: A User's Guide". Cambridge England. New York: Cambridge University Press.

Miller, J. (1998). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 14(3),* 539-543.

Metropolis, M. & Ulam, S. (1949). The Monte Carlo Method. *Journal of American Statistical Association*, 44(247), 335-341.

Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods, 45(3),* 792-812. 33

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance, 37(1),* 58.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6(10),* 421-425.

Posner, M., Snyder, C. & Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General, 109(2).* 160-174. DOI:http://dx.doi.org/10.1037/0096-3445.109.2.160

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

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114(3),* 510-532.

Tape, T.G. Interpreting Diagnostic Tests [Website]. Retrieved from gim.unmc.edu/roc2.htm

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123(1),* 34-80.

Van Breukelen, G. J. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70(2),* 359-376.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7(3),* 424-465.

Veksler, V.D., Myers, C.W., & Gluck, K.A. (2015). Model flexibility analysis. *Psychological Review, 122(4),* 755-769.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision research, 50(14),* 1304-1311.

Wright, D.B. and London, K. 2009. Multilevel modeling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology, 62,* 439-456.

Appendix A

SAS Code

```
*****************************************************************;
* Angela Crumer Project: Continuous Simulation (All Responses) *;
* Coder: Nick Bloedow (Statistical Consulting Lab @ KSU)       *;
* Date: March 30, 2016                                         *;
*****************************************************************;
/*
Notes: Extracts key information from the Simulation Study into a exportable csv file with proper labels
*/
%macro create(mu,j,var);
data gamma;
do subID = 1 to &j; /* Number of Subjects */
do submu=rand('normal',&mu,&var); /* Subject Mean */
do subint=rand('normal',6.8,1); /* Subject Intercept */
        do trials=1 to 10 by 1;  /* Number of trials/replicates per subject*/
        condmean=1+ 10 * ((trials-1)/9); /*Condition Mean given trial # with slope=5*/
        response=(subint+submu*condmean)*rand('GAMMA',1.5);  /* Generation of Response given
Mean/Intercept/Condition Mean */
   log_response=log(response);  /* Log of Response */
output;
end;
 end;
end;
end;
keep subID trials condmean response log_response submu subint;
run;

proc sort data=gamma;
  by subid trials;
run;
```

```
proc means data=gamma mean std;
  var response;
  by subid;
  output out=gamma_sum(drop=_TYPE_);
run;

proc transpose data=gamma_sum out=trans_sum(drop=_name_);
  where _STAT_="MEAN" | _STAT_="STD";
  id _STAT_;
  var response;
  by subid;
run;

proc sort data=gamma;
  by subID trials ;
run;

data gamma;
  merge gamma trans_sum;
  by subID;
run;

data gamma;
  set gamma;
  if abs(Response-MEAN)<(2.5*STD) then trunc_response=response ; else trunc_response=.;
  drop MEAN STD;
run;

proc means data=gamma;
  var trunc_response;
  output out=miss_trunc(drop=_TYPE_) NMISS=;
```

```
run;

proc transpose data=miss_trunc out=trans_miss_trunc(drop=_name_) prefix=Miss_Cond;
  var trunc_response;
run;

proc append base=miss_final data=trans_miss_trunc;
run;

dm "log; clear; ";

/*proc sgplot data=gamma;
  scatter x=Condmean y=response / group=SubID;
  series  x=Condmean y=response / group=SubID;
  xaxis label="Condition Mean";
run;

proc sgpanel data=gamma;
  panelby SubID / columns=5 rows=2;
  scatter x=Condmean y=response / group=SubID;
  series  x=Condmean y=response / group=SubID;
  colaxis label="Condition Mean";
run;

proc sgplot data=gamma;
  scatter x=Condmean y=log_response / group=SubID;
  series  x=Condmean y=log_response / group=SubID;
  xaxis label="Condition Mean";
run;

proc sgpanel data=gamma;
```

```
  panelby SubID / columns=5 rows=2;
  scatter x=Condmean y=log_response / group=SubID;
  series  x=Condmean y=log_response / group=SubID;
  colaxis label="Condition Mean";
run;

proc sgplot data=gamma;
  scatter x=Condmean y=trunc_response / group=SubID;
  series  x=Condmean y=trunc_response / group=SubID;
  xaxis label="Condition Mean";
run;

proc sgpanel data=gamma;
  panelby SubID / columns=5 rows=2;
  scatter x=Condmean y=trunc_response / group=SubID;
  series  x=Condmean y=trunc_response / group=SubID;
  colaxis label="Condition Mean";
run;
*/
%mend create;
%macro analyze1(resp);

title3 "Response: &resp";
title4 "Number of Subjects: &j";
title5 "Mu: &mu";

data descrip;
  length Response $15;
  Response="&resp";
  Iteration=&k;
  n=&j;
```

```
  mu=&mu;
run;

proc glimmix data=gamma abspconv=0.00001 ic=q ;
  class subID;
  model &resp=trials/dist=gamma solution ;
  random subID; *subID*trials;
  ods output Tests3=test_gamma FitStatistics=fit_gamma ParameterEstimates=par_gamma;
  title6 "Gamma Analysis";
run;

data test_gamma;
  set test_gamma(rename=(NumDF=Gamma_NumDF DenDF=Gamma_DenDF FValue=Gamma_FValue
ProbF=Gamma_ProbF));
  drop Effect;
run;

data fit_names_gamma;
  length Statistic $25;
    Statistic="Gamma_-2 Res Log Likelihood";
          output;
          Statistic="Gamma_AIC";
          output;
          Statistic="Gamma_AICC";
          output;
          Statistic="Gamma_BIC";
          output;
          Statistic="Gamma_CAIC";
          output;
          Statistic="Gamma_HQIC";
          output;
          Statistic="Gamma_Generalized Chi-Square";
```

```
        output;
        Statistic="Gamma_Gener. Chi-Square_DF";
        output;
run;

data fit_gamma;
  merge fit_names_gamma fit_gamma;
run;

data fit_gamma;
  set fit_gamma;
  where Statistic="Gamma_AIC" | Statistic="Gamma_BIC";
  drop Descr;
run;

proc transpose data=fit_gamma out=trans_fit_gamma(drop=_name_);
  id Statistic;
  var Value;
run;

data par_int_gamma;
  set par_gamma(rename=(Estimate=Gamma_Est_int StdErr=Gamma_StdErr_int DF=Gamma_DF_int
tValue=Gamma_tValue_int Probt=Gamma_Probt_int));
  where Effect="Intercept";
  drop  Effect;
run;

data par_trials_gamma;
  set par_gamma(rename=(Estimate=Gamma_Est_trials StdErr=Gamma_StdErr_trials
DF=Gamma_DF_trials tValue=Gamma_tValue_trials Probt=Gamma_Probt_trials));
  where Effect="trials";
  drop Effect;
```

```
run;

data final_gamma;
  merge test_gamma trans_fit_gamma par_int_gamma par_trials_gamma;
run;

dm "log; clear; ";
proc glimmix data=gamma abspconv=0.00001;
  class subID;
  model &resp=trials/dist=normal solution ;
  random subID; *subID*trials;
  ods output Tests3=test_normal FitStatistics=fit_normal ParameterEstimates=par_normal;
  title6 "Normal Analysis";
run;

data test_normal;
  set test_normal(rename=(NumDF=Normal_NumDF DenDF=Normal_DenDF FValue=Normal_FValue
ProbF=Normal_ProbF));
  drop Effect;
run;

data fit_names_normal;
  length Statistic $25;
    Statistic="Normal_-2 Res Log Likelihood";
            output;
            Statistic="Normal_AIC";
            output;
            Statistic="Normal_AICC";
            output;
            Statistic="Normal_BIC";
            output;
            Statistic="Normal_CAIC";
```

```
        output;
        Statistic="Normal_HQIC";
        output;
        Statistic="Normal_Generalized Chi-Square";
        output;
        Statistic="Normal_Gener. Chi-Square_DF";
        output;
run;

data fit_normal;
  merge fit_names_normal fit_normal;
run;

data fit_normal;
  set fit_normal;
  where Statistic="Normal_AIC" | Statistic="Normal_BIC";
  drop Descr;
run;

proc transpose data=fit_normal out=trans_fit_normal(drop=_name_);
  id Statistic;
  var Value;
run;

data par_int_normal;
  set par_normal(rename=(Estimate=Normal_Est_int StdErr=Normal_StdErr_int DF=Normal_DF_int
tValue=Normal_tValue_int Probt=Normal_Probt_int));
  where Effect="Intercept";
  drop  Effect;run;

data par_trials_normal;
```

```
  set par_normal(rename=(Estimate=Normal_Est_trials StdErr=Normal_StdErr_trials
DF=Normal_DF_trials tValue=Normal_tValue_trials Probt=Normal_Probt_trials));
  where Effect="trials";
  drop Effect;run;

data final_normal;
  merge test_normal trans_fit_normal par_int_normal par_trials_normal;run;

data final_combined;
  merge descrip final_gamma final_normal;run;

proc append base=final1 data=final_combined;run;

*Deletes temporary data sets used for creating the final combined results dataset for each simulation;
proc delete data=Work.descrip; run;
proc delete data=Work.test_gamma; run;
proc delete data=Work.fit_names_gamma; run;
proc delete data=Work.fit_gamma; run;
proc delete data=Work.trans_fit_gamma; run;
proc delete data=Work.par_gamma; run;
proc delete data=Work.par_int_gamma; run;
proc delete data=Work.par_trials_gamma; run;
proc delete data=Work.final_gamma; run;
proc delete data=Work.test_normal; run;
proc delete data=Work.fit_names_normal; run;
proc delete data=Work.fit_normal; run;
proc delete data=Work.trans_fit_normal; run;
proc delete data=Work.par_normal; run;
proc delete data=Work.par_int_normal; run;
proc delete data=Work.par_trials_normal; run;
proc delete data=Work.final_normal; run;
proc delete data=Work.final_combined; run;
```

```
%mend analyze1;
dm "log; clear; ";

*ods rtf file = "C:\Users\Angela\Desktop\PhD Work\Simulation Stuff\Continuous Simulation Study
Project (01_27_17).doc";

*title "Continuous Simulation Study (Angela Crumer)";

%let max=50;
%macro simulation(mu,j,var);
  %do k = 1 %to &max %by 1;
    title2 "Iteration: &k";
    %create(&mu,&j,&var);
    %analyze1(response);
    %analyze1(log_response);
    %analyze1(trunc_response);
    *proc delete data=Work.gamma; run;
  %end;
%mend simulation;

/* Simulation Settings
    max: 1000
    mu: 0, 0.25, 0.50, & 1
    var: 0.06, 0.06, 0.11, & 0.22
    j: 30, 60, & 120
*/
* Values to simulate a small sample size;
  %simulation(0,30,0.06); *Values to simulate no effect;
  %simulation(.25,30,0.06); *Values to simulate a small effect;
  %simulation(.5,30,0.11);  *Values to simulate a medium effect;
  %simulation(1,30,0.22); *Values to simulate a large effect;
```

```
* Values to simulate aforementioned effect sizes for a medium sample size;
  %simulation(0,60,0.06);
  %simulation(.25,60,0.06);
  %simulation(.5,60,0.11);
  %simulation(1,60,0.22);

* Values to simulate aforementioned effect sizes for a large sample size;
  %simulation(0,120,0.06);
  %simulation(.25,120,0.06);
  %simulation(.5,120,0.11);
  %simulation(1,120,0.22);
dm "log; clear; ";
  data Response;
    set Final1;
            where Response="response";
            drop Response;  run;

  proc export data=Response outfile="C:\Users\Angela\Desktop\PhD Work\Simulation Stuff\Data from
Continuous Simulations with 10 trials\RAW\Continuous Simulation 10_19 RAW.csv" dbms=csv replace;
  run;

  data Log_Response;
    set Final1;
            where Response="log_response";
            drop Response;  run;

  proc export data=Log_Response outfile="C:\Users\Angela\Desktop\PhD Work\Simulation Stuff\Data
from Continuous Simulations with 10 trials\LOG\Continuous Simulation 10_19 LOG.csv" dbms=csv
replace;
  run;
  data Trunc_Response;
```

```
  set Final1;
        where Response="trunc_response";
        drop Response;
 run;
 data Trunc_Response;
  merge miss_final Trunc_Response;  run;


 proc export data=Trunc_Response outfile="C:\Users\Angela\Desktop\PhD Work\Simulation Stuff\Data
from Continuous Simulations with 10 trials\TRUNC\Continuous Simulation 10_19 TRUNC.csv"
dbms=csv replace;
 run;
*ods rtf close;
```

Appendix B

Additional Results

SDT Theory

In addition to the significant effects involving technique, there were significant main effects of the number of trials per subject and the sample size for both types of predictors (Table 5). There was also a significant interaction effect involving the number of trials per subject, and sample size for both parameters for continuous predictors (Table 6).

Interaction effect.

When continuous predictors were used, there was a significant interaction between the number of trials per subject and the sample size. The difference between 5 and 10 trials is slightly less for larger sample sizes (Fig B1).

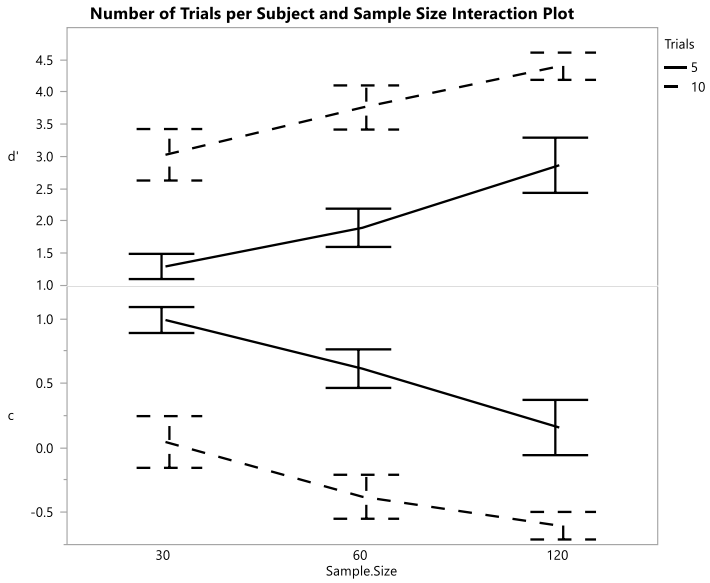

Figure B1 Number of trials per subject and sample size interaction plot. This figure is the interaction plot for number of trials per subject and sample size for continuous predictors

Main effects.

Although some main effects are qualified by the significant interactions, they are still very apparent in the data. All main effects significantly impacted *d'* and *c* for continuous and discrete predictors (Table 5).

There was a significant effect of the number of trials per subject on both the discriminability measure and the criterion value for both discrete and continuous predictors (Table 5). Marginal analyses indicated that more trials resulted in higher discriminability and a more neutral criterion value (Fig B2). This result supports the first major hypothesis that increasing the number of observations will increase *d'* values and result in more neutral *c* values. Data created using 10 trials per subject resulted in more variability in both *d'* and *c* for discrete predictors and skewed the distribution when continuous predictors were used (Fig B2).

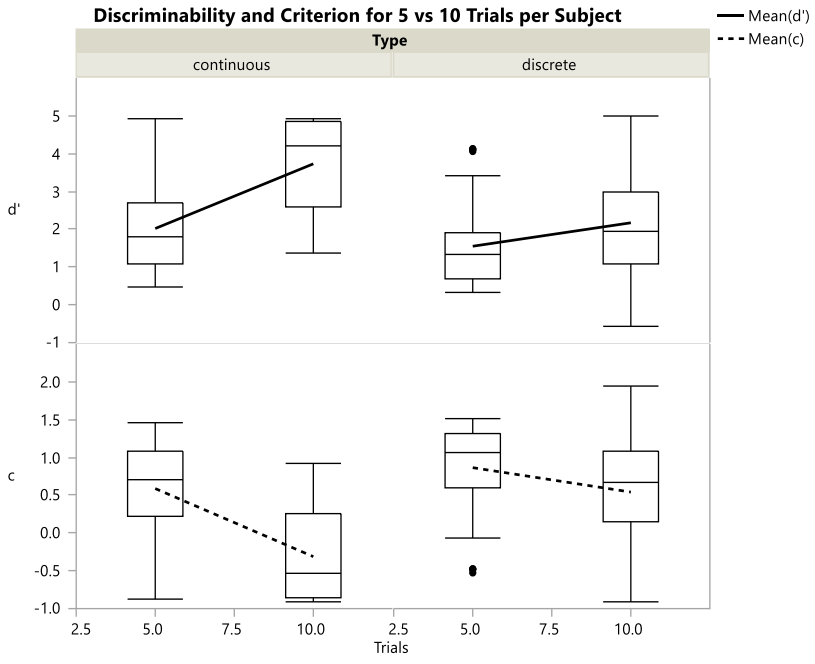

Figure B2 Discriminability and criterion for 5 v 10 trials per subject. This figure is a plot of the main effect for the number of trials per subject on *d'* and *c* for continuous and discrete predictors.

There was a significant effect of the number of subjects on both the discriminability measure and the criterion value (Table 5). More subjects resulted in higher discriminability and a more neutral criterion for both discrete and continuous predictors. Fewer subjects also resulted in more variability in both *d'* and *c* calculations for discrete predictors but did not affect variability in continuous predictors (Fig B3).

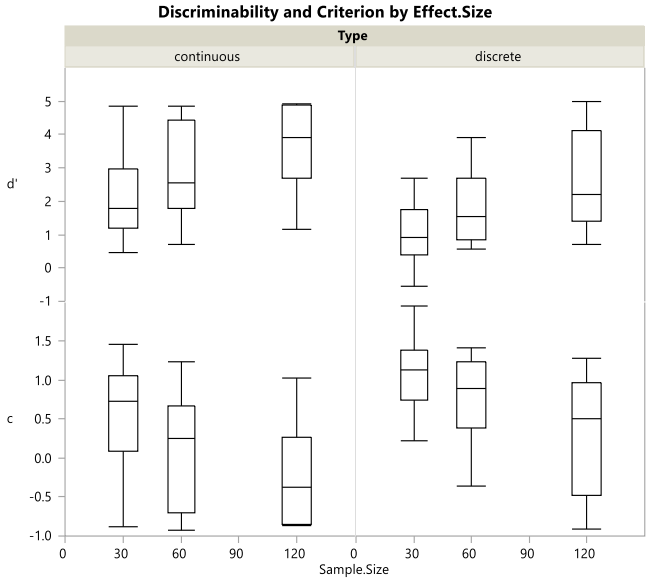

Figure B3 Discriminability and criterion by sample size. Main effect plot for sample size on *d'* and c for continuous and discrete predictors.

To control the experimentwise error rate at α=0.05, Tukey's honestly significant difference correction was used in comparing the least squares means for each sample size. This indicates that all three sample sizes were significantly different from one another for both *d'* and c for discrete predictors (Table B1). As predicted, larger sample sizes increased *d'* and decreased c.

Table B1

Post hoc means comparisons for effect size for d' and c for discrete predictors. Sizes with the same letter are not significantly different from each other.

For *d'* Effect Size		Least Sq Mean	For c Effect Size		Least Sq Mean
120	A	3.962	30	A	0.773
60	B	2.606	60	B	0.412
30	C	1.706	120	C	-0.404

To control the experimentwise error rate at α=0.05, Tukey's honestly significant difference correction was used in comparing the least squares means for each sample size. This indicates that all three sample sizes were significantly different from one another for both *d'* and c for continuous predictors (Table B2).

Table B2

Post hoc means comparisons for effect size for d' and c for continuous predictors. Sizes with the same letter are not significantly different from each other.

For *d'* Effect Size		Least Sq Mean	For c Effect Size		Least Sq Mean
120	A	3.633	30	A	0.519
60	B	2.828	60	B	0.118
30	B	2.162	120	B	-0.223

Parameter Recovery

When estimating the parameters, there were also several significant interactions between the number of trials per subject, sample size, and effect size (Table 12).

General results.

In general, larger sample sizes, larger numbers of trials per subject, and smaller effect sizes (Fig B4) result in less biased and more accurate intercept and difference (or slope) estimates for both continuous and discrete predictors, with a few exceptions. This supports the prediction that increasing the number of observations will result in more accurate parameter

estimates. The effects of these predictors are sometimes mitigated by different techniques and will be discussed later.

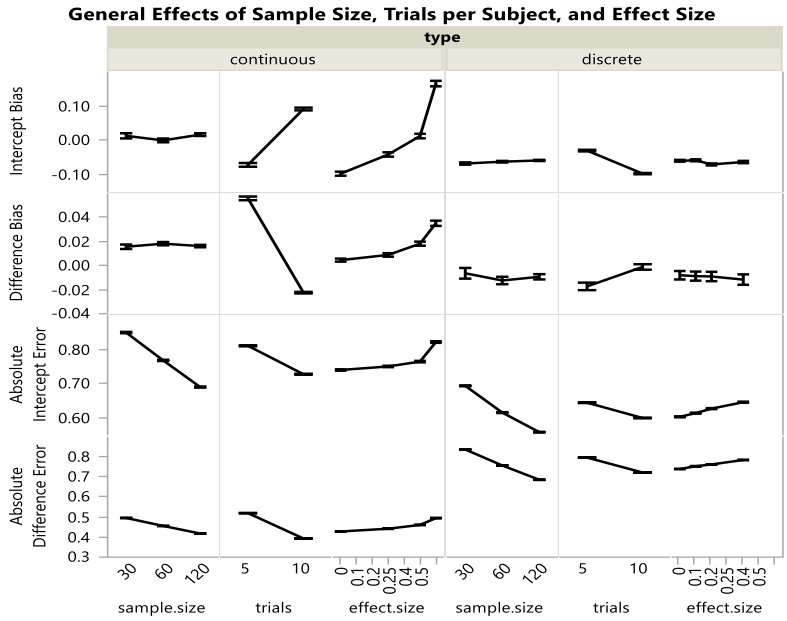

Figure B4 General effects of sample size, trials per subject, and effect size. This graph shows the impact of sample size, trials per subject, and effect size on estimation bias and accuracy for continuous and discrete predictors. Each error bar is constructed using 1 standard error from the mean.

Effects of each technique.

Once it was determined that the techniques had differing effects on each of the linear-scale dependent variables, separate analyses were conducted for each technique. These were three-way factorial models. For these analyses there were four models:

$$Raw\ Error\ of\ the\ Intercept = OM + [trials\ per\ subject]_i + [sample\ size]_j + [effect\ size]_k +$$

to assess the impact of trials per subject, sample size, and effect size on the intercept estimates for each technique;

$$Raw\ Error\ of\ the\ Difference = OM + [trials\ per\ subject]_i + [sample\ size]_j + [effect\ size]_k$$

to assess the impact of trials per subject, sample size, and effect size when estimating the difference (for discrete predictors) or slope (for continuous predictors);

$$\forall(|Intercept\ Error|) = OM + [trials\ per\ subject]_{\downarrow}i + [sample\ size]_{\downarrow}j + [effect\ size]_{\downarrow}k + [trial$$

to assess the impact of trials per subject, sample size, and effect size on the amount of bias in the intercept estimates for each technique;

$$\forall(|Difference\ Error|) = OM + [trials\ per\ subject]_{\downarrow}i + [sample\ size]_{\downarrow}j + [effect\ size]_{\downarrow}k + [tr$$

to assess the impact of trials per subject, sample size, and effect size on the amount of bias when estimating the difference (for discrete predictors) or slope (for continuous predictors).

Linear-scale techniques.

Ignoring distributional assumptions.

Ignoring the distributional assumptions did not create estimation bias in the intercept when discrete predictors were used. When continuous predictors were used, absent of other factors, ignoring distributional assumptions caused significant underestimation, *t=-7.23*, *p<0.001*, large effects to be significantly lower than other effect sizes, *F(3,3)=9.14*, *p<0.001*, and 5 trials led to underestimation while 10 trials was basically unbiased, *F(1,1)=42.72*, *p<0.001*. The intercept estimates were significantly inaccurate for continuous predictors, *F(1,23)=153.81*, *p<0.001*. Ignoring distributional assumptions caused significantly inaccurate estimates of the intercept for continuous, *t=508.85*, *p<0.0001* and discrete, *t=518.14*, *p<0.0001*, predictors. There was a significant effect of the number of trials per subject for both continuous, *F(1,1)=875.66*, *p<0.001*, and discrete, *F(1,1)=602.61*, *p<0.001*, predictors. Increasing the number of trials per subject increased accuracy for both types of predictors. There was also a significant effect of the sample size for both discrete and continuous predictors, *F(2,2)=1213.55*, *p<0.001* and *F(2,2)=1073.06*, *p<0.001*, respectively. Increasing the sample size increased accuracy for both types of predictors. There was also a significant impact of effect size for both discrete and continuous predictors, *F(3,3)=85.52*, *p<0.001* and *F(3,3)=162.32*, *p<0.001*, respectively. Increasing the effect size decreased the accuracy of the intercept estimate for both types of predictors. For continuous predictors, there was a significant interaction between the

sample size and effect size, *F(6,6)=2.85, p=0.0088*. Small effect sizes behaved slightly different than other effect sizes when the sample size was sixty, and for discrete predictors, there was a significant interaction between the number of trials per subject and sample size, *F(2,2)=6.67, p=0.0013*. Both effects were very minimal.

There was no bias in estimating the difference for discrete predictors. In fact, the overall model was non-significant. For continuous predictors, ignoring distributional assumptions caused overestimation of the difference parameter, *t=21.33, p<0.001*, and resulted in significant impacts of the number of trials per subject, *F(1,1)=440.34, p<0.001*, and effect size, *F(3,3)=86.11, p<0.001*. Fewer trials per subject led to overestimation of the difference. In fact, there was virtually no estimation bias for 10 trials per subject (*M=0.001, SD=0.191*) for continuous predictors. Increasing the effect size increased the estimation bias only for continuous predictors. However, both effects were qualified by a significant interaction between the number of trials per subject and the effect size for continuous predictors, *F(3,3)=61.76, p<0.001*. Bias in estimating the difference parameter increased with effect size for 5 trials but remained consistently unbiased for 10 trials (Fig B5).

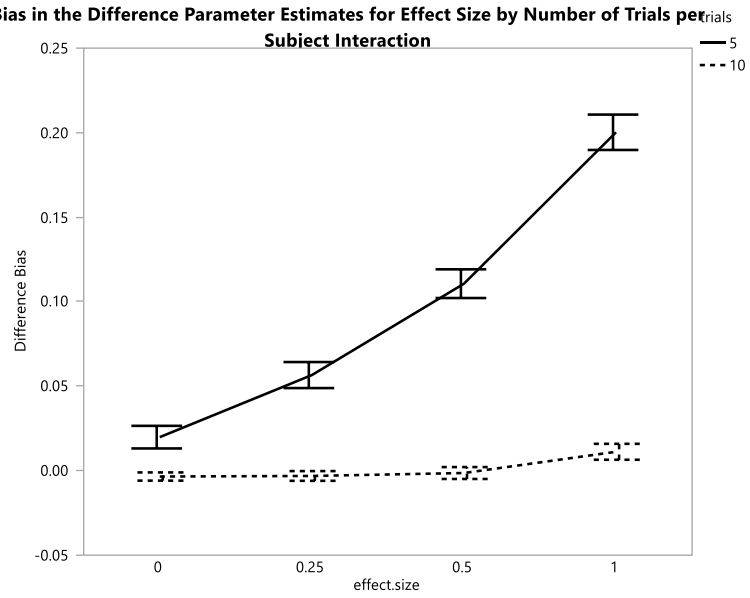

Figure B5 Bias in the difference parameter estimates for effect size by number of trials per subject interaction. This figure shows the interaction plot for effect size and number of trials per subject for continuous predictors when distributional assumptions are ignored. Each error bar is constructed using 1 standard error from the mean.

Ignoring the distributional assumptions created inaccurate estimates of the difference for both continuous predictors, $t=573.63$, $p<0.001$. It further resulted in significant impacts of all main effects and several significant interactions. More observations and smaller effect sizes led to more accurate difference estimates for both types of predictors. The number of trials per subject interacted with the sample size. However, this difference is minimal and likely caused by the small mean square error. The notable interaction is the sample size by effect size interaction. All effect sizes behave similarly when increasing the sample size from 60 to 120, but they differ in their accuracy when increasing from 30 to 60 (Fig B6). For continuous predictors, there was also a significant three-way interaction, $F(6,6)=2.23$, $p=0.0375$, but is uninterpretable and likely caused by an overpowered test.

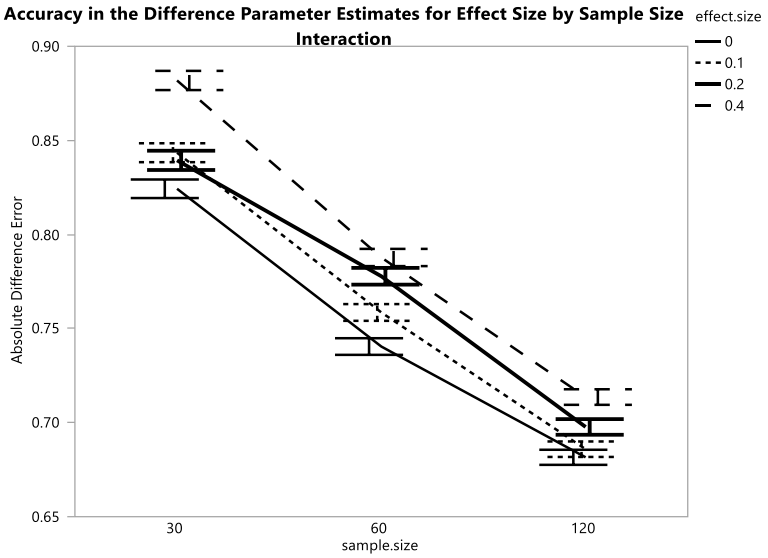

Figure B6 Accuracy in the difference parameter estimates for effect size by sample size interaction. This figure shows the interaction plot for sample size and effect size for discrete predictors when distributional assumptions are ignored. Each error bar is constructed using 1 standard error from the mean.

Aggregation.

Aggregating the data did not lead to bias in the intercept or difference parameter estimates. Both omnibus models were non-significant.

Aggregation did, however, cause inaccurate estimates for both the intercept, *t=512.1, p<0.001* and the difference, *t=588.49, p<0.001*. Furthermore, when data were aggregated, there were significant main effects of each predictor in the models, as well as significant two-way interactions. Both parameter estimates were more accurate when there were more trials per subject, larger sample sizes, and smaller effect sizes. In both the intercept and difference parameter estimates, there was a significant interaction between the number of trials per subject and sample size, *F(1,1)=5.61, p=0.0037* and *F(2,2)=4.77, p=0.0085*, respectively. The pattern was the same for both: sample sizes of 60 and 120 improve accuracy at the same rate as the number of trials increases, but the accuracy improves slightly more for a sample size of thirty. There was also a significant interaction between the sample size and effect size for the difference

parameter. Larger sample sizes were less affected by effect size than smaller sample sizes. The accuracy of the difference parameter estimate for the smallest sample size (n=30) varied much more than the other sample sizes at different effect sizes (Fig B7).

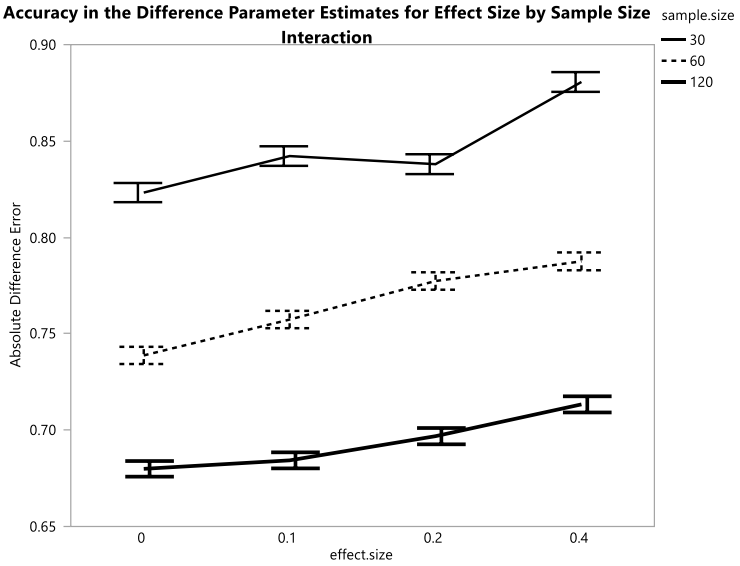

Figure B7 Accuracy in the difference parameter estimates for effect size by sample size interaction. This figure shows the interaction plot for effect size and sample size when data are aggregated. Each error bar is constructed using 1 standard error from the mean.

Truncation.

Truncating the data produced varying results, depending on whether the predictors were continuous or discrete. Either way, truncating the data significantly biased the intercept parameter estimates for both discrete, $t=-39.02$, $p<0.001$ and continuous $F(1,23)=-3.42$, $p=0.0006$, predictors. However, these results are better understood by looking at the main effect of the number of trials per subject. When there were only 5 trials per subject, truncation did not occur. With continuous predictors, there was significant underestimation of the intercept parameter estimate when truncation did not occur and virtually no estimation bias when truncation did occur. However, the negative estimation bias is minimal and is likely due to the small mean square error at 5 trials ($M=-0.11$, $SD=1.42$) and 10 trials ($M=0.06$, $SD=1.01$). When

the predictors were discrete, the opposite was true. There was significant underestimation when truncation did occur (*M=-0.3, SD=0.49*) and no estimation bias when it did not occur (*M=0.01, SD=0.64*). The difference between the number of trials in this case is more apparent (Fig B8).

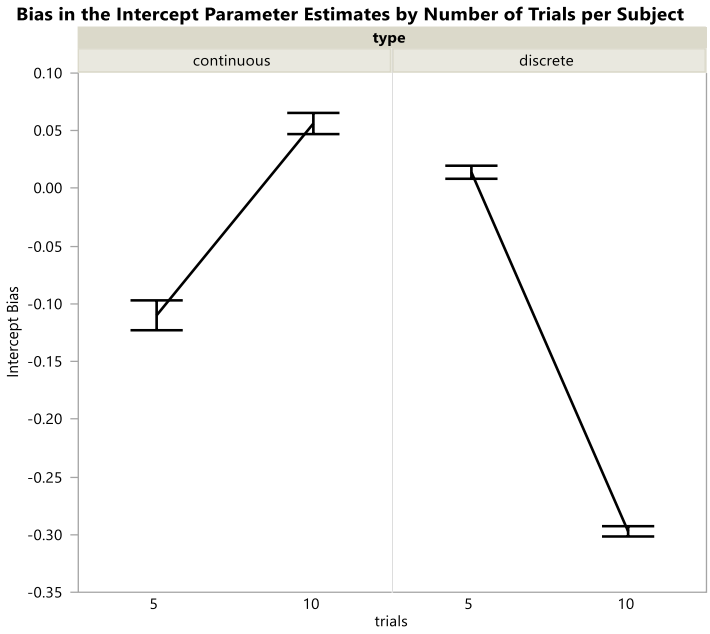

Figure B8 Bias in the intercept parameter estimates by number of trials per subject. This figure shows the error in the intercept parameter estimates by the number of trials per subject when data are truncated. Each error bar is constructed using 1 standard error from the mean.

This effect was qualified by a significant two-way interaction with the effect size. This is to be expected since the number of trials determines truncation. For discrete predictors, this interaction, *F(3,3)=3.16, p=0.0237*, showed virtually no estimation bias when truncation did not occur and increasingly negative estimation bias as effect size increased when truncation did occur. For continuous predictors, this interaction, *F(3,3)=122.43, p<0.001*, showed an interesting effect of truncation as effect size increased. When truncation did not occur, increasing effect size mildly underestimated intercept parameter estimate. When truncation did occur,

parameter estimates were underestimated for small effect sizes, but quickly became greatly overestimated as the effect size increased (Fig B9).

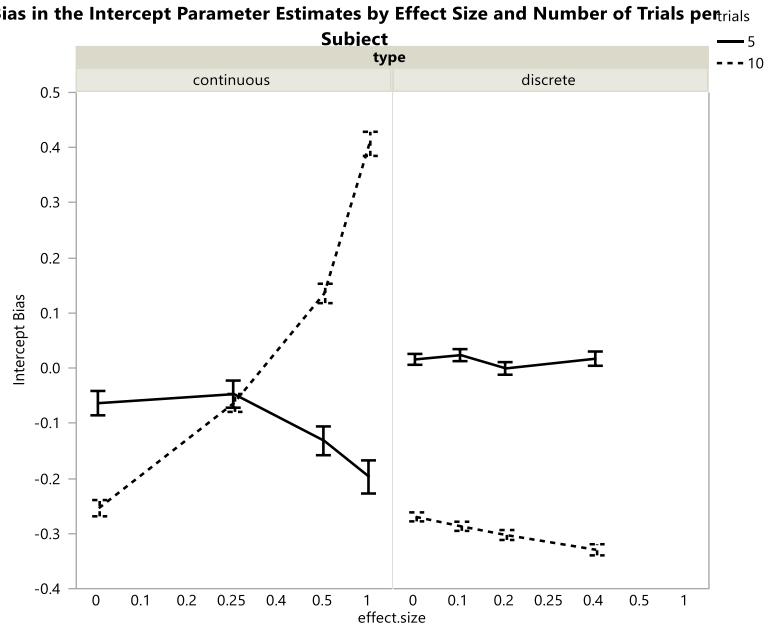

Figure B9 Bias in the intercept parameter estimates by effect size and number of trials per subject. This figure shows the interaction plot for effect size and number of trials per subject when data are truncated. Each error bar is constructed using 1 standard error from the mean.

There was also a significant two-way interaction between the number of trials per subject and the sample size, $F(2,2)=3.61$, $p=0.0271$, when continuous predictors were used. Estimates were underestimated when truncation did not occur, overestimated when truncation did occur, and the overestimation increased as the sample size increased (Fig B10).

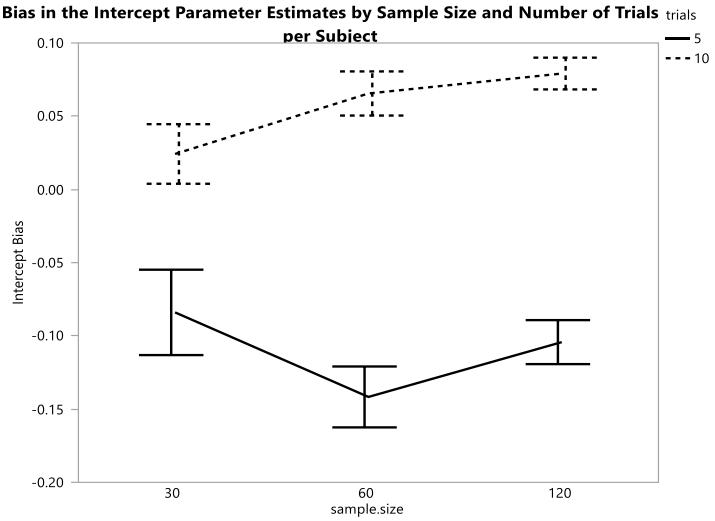

Figure B10 Bias in the intercept parameter estimates by sample size and number of trials per subject. This figure shows the interaction plot for sample size and the number of trials per subject for the bias in the intercept parameter estimates with continuous predictors when data are truncated. Each error bar is constructed using 1 standard error from the mean.

Given the amounts of estimation bias in the intercept parameter estimates, it is no surprise that there is significant error in the estimates as well, for both continuous, $t=510.25$, $p<0.0001$, and discrete, $t=528.77$, $p<0.001$, predictors. Estimates were slightly more accurate when truncation occurred for both types of predictors, although the difference was minimal. Following the overall pattern, for both types of predictors, estimates were more accurate for larger sample sizes and smaller effect sizes. The more notable results regarding the accuracy of the intercept estimates are the two-way interactions. For discrete predictors, only the number of trials per subject by sample size interaction was significant, whereas all two-way interactions were significant for continuous predictors. For both continuous and discrete predictors, as sample size increased, accuracy increased more quickly when truncation did not occur than when it did, $F(2,2)=4.2$, $p=0.015$ and $F(2,2)=46.6$, $p<0.001$, respectively. For continuous predictors, as effect size increased, accuracy decreased more quickly when truncation occurred than when it

did not occur, *F(3,3)=5.55, p=0.0008*. There was also a significant interaction between sample size and effect size for continuous predictors, *F(6,6)=2.61, p=0.0159.* For the smallest sample size, accuracy decreased more quickly than the others when increasing from no effect to a small effect.

There was no overall effect of truncation on the difference parameter estimation bias. However, the effects of truncation are explained by the main effect of the number of trials per subject. For both types of predictors, there were approximately equal amounts of estimation bias with and without truncation, but in different directions. Additionally, data values were more spread out when truncation did not occur. Truncation was mitigated by an interaction with effect size. For continuous predictors, increasing effect size caused increasing overestimation when truncation did not occur and increasing underestimation when truncation did occur. For discrete predictors, increasing effect size did not affect the parameter estimates when truncation did not occur, but caused increasing overestimation when truncation did occur (Fig B11).

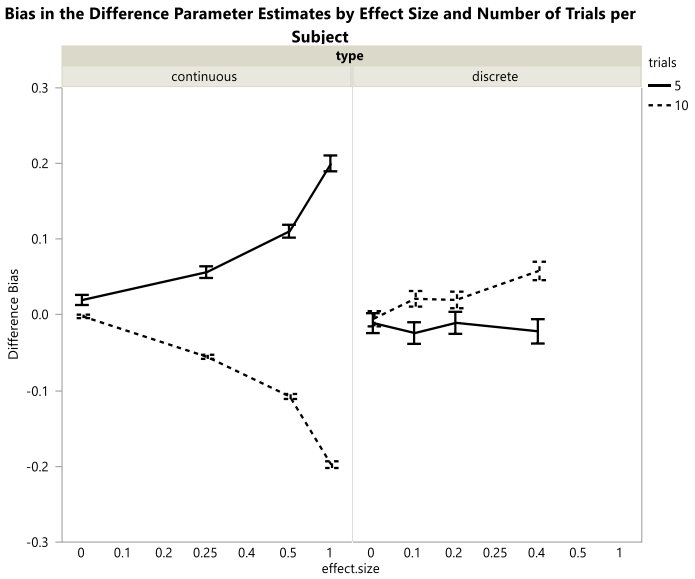

Figure B11 Bias in the difference parameter estimates by effect size and number of trials per subject. This graph shows the interaction plot for effect size and number of trials per subject for

bias in the difference parameter estimates when data are truncated.
Each error bar is constructed using 1 standard error from the mean.

In light of the amounts of estimation bias seen in the difference parameter estimates, it is no surprise that there are significant amounts of error in them as well. The difference parameter estimates were more accurate when truncation occurred than when it did not for both continuous, $F(1,23)=4590.46$, $p<0.0001$, and discrete, $F(1,23)=732.73$, $p<0.0001$, predictors. Larger sample sizes and smaller effect sizes resulted in more accurate estimates for both types of predictors. Accuracy increased more quickly as sample size increased when truncation did not occur than when it did for both continuous, $F(2,2)=44.94$, $p<0.0001$, and discrete, $F(2,2)=4.35$, $p=0.013$, predictors. For continuous predictors, accuracy decreased more quickly as effect size increased when truncation occurred than when it did not, $F(3,3)=44.59$, $p<0.0001$. There was also a significant three-way interaction, $F(6,6)=4.11$, $p=0.0004$. This is mostly uninterpretable and is likely due to the small mean square error ($MSE=0.04$).

Log-scale techniques.

Transformation.

Transforming the data with a logarithmic transformation using continuous predictors resulted in a significant impact of the number of trials per subject, $F(1,1)=400.34$, $p<0.001$, and effect size, $F(1,1)=108.2$, $p<0.001$. However, both were qualified by a significant interaction, $F(3,3)=86.9$, $p<0.001$. The intercept estimates were significantly underestimated and more dispersed for 5 trials and virtually unbiased for 10 trials. There was generally no estimation bias when there was no effect or a small effect size, but increasing underestimation as the effect size increased. These effects interacted with each other such that 10 trials showed virtually no estimation bias for all effect sizes, but larger effect sizes showed more underestimation than smaller effect sizes when there were only 5 trials per subject (Fig B12).

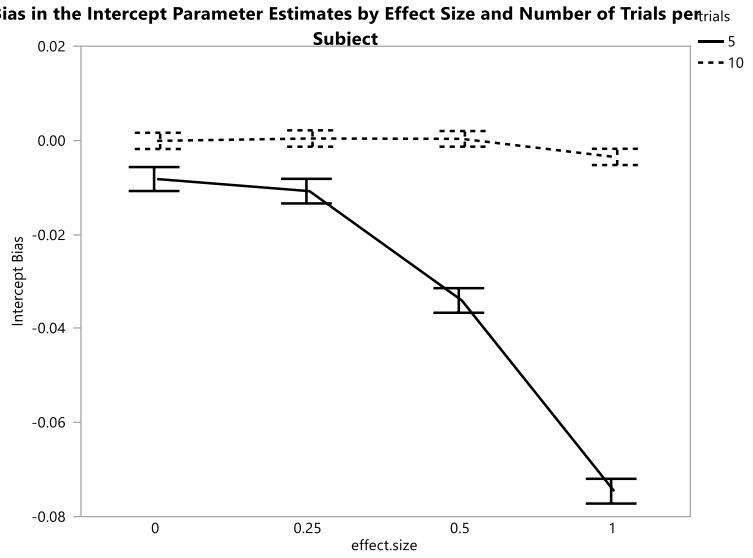

Figure B12 Bias in the intercept parameter estimates by effect size and number of trials per subject. This graph shows the interaction plot of effect size and the number of trials per subject for bias in the intercept parameter estimates when data are transformed. Each error bar is constructed using 1 standard error from the mean.

When discrete predictors were used, the only significant impact on the intercept parameter estimates was the effect size, $F(3,3)=5.39$, $p=0.0011$. The medium effect size had significantly lower estimates than the small and large effect sizes (Table B3).

Table B3

Post hoc means comparisons of effect size for bias in the intercept estimate for discrete predictors when data are transformed. Techniques with the same letter are not significantly different from each other.

		Mean	N	Effect Size
	A	-0.0049	6150	Small
	A	-0.0059	6150	Large
B	A	-0.0075	6150	None
B		-0.0087	6150	Medium

Given the amount of underestimation when the log transformation was used, it was not surprising to find significant error in the intercept parameter estimates for both the continuous, $t=508.94$, $p<0.001$, and the discrete, $t=516.18$, $p<0.001$, predictors. Both continuous and discrete predictors had significant impacts of the number of trials per subject, $F(1,1)=1505.14$, $p<0.001$ and $F(1,1)=592.05$, $p<0.001$, and sample size, $F(2,2)=936.7$, $p<0.001$ and $F(2,2)=1183.36$, $p<0.001$, respectively. When continuous predictors were used, these effects were qualified by a significant interaction term, $F(6,6)=3.05$, $p=0.0475$. More trials per subject improves parameter estimates, but this improvement is greater for the smaller sample size than for the larger ones. Also when continuous predictors were used effect size impacted accuracy in the difference estimates, $F(3,3)=12.25$, $p<0.001$, and significantly interacted with the number of trials per subject, $F(3,3)=12.46$, $p<0.001$. Large effects were significantly less accurate than other effect sizes. Increasing the number of trials per subject improved accuracy in the intercept parameter estimate.

Applying a logarithmic transformation to the data resulted in biased estimates of the difference parameter. This estimation bias depended on the type of predictors used. When using continuous predictors, transforming the data resulted in significant overestimation, $t=34.81$, $p<0.001$, while using discrete predictor resulted in significant underestimation, $t=-9.12$, $p<0.001$. The significance found when using continuous predictors was likely due to the small amount of variability and extremely small mean square error ($MSE=0.001$).

There was a significant impact of effect size for both discrete, $F(3,3)=14.22$, $p<0.001$, and continuous, $F(3,3)=283.53$, $p<0.001$, predictors. The latter was also qualified by a significant effect of the number of trials per subject, $F(1,1)=1276.02$, $p<0.001$, and interaction, $F(3,3)=256.28$, $p<0.001$. As effect size increased, estimation bias significantly decreased. For

continuous predictors, increasing the number of trials per subject decreased the variability in the estimates and basically eliminated the estimation bias. There is a significant difference between effect sizes when there were 5 trials per subject but the difference was negligible when there were 10 trials (Fig B13).

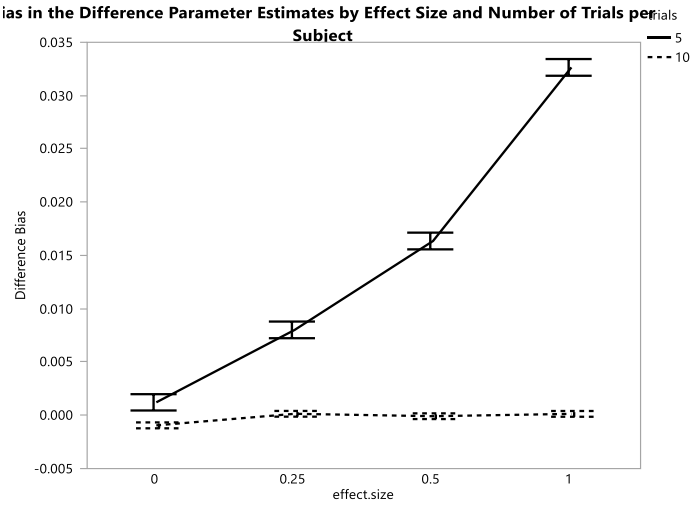

Figure B13 Bias in the difference parameter estimates by effect size and number of trials per subject. This graph shows the interaction plot for effect size and the number of trials per subject for the bias in the intercept parameter estimates with continuous predictors when data are transformed. Each error bar is constructed using 1 standard error from the mean.

There were many significant effects and interactions on the accuracy of the difference parameter estimates. Transforming the data resulted in significantly inaccurate estimates for both the discrete, $t=592.95$, $p<0.001$, and the continuous, $t=578.74$, $p<0.001$, predictors. All main effects were significant for both types of predictors. In general, more observations and smaller effects produced more accurate estimates of the difference. These individual effects are best understood through their interactions. There was a significant interaction between the number of trials per subject and sample size for continuous predictors, $F(3,3)=3.35$, $p=0.0351$. Smaller sample sizes improve with an increase in the number of trials per subject ever so slightly more than the larger sample sizes. This effect is likely significant because of the tiny mean square

errors. When continuous predictors were used, the large effect size was affected by increasing the number of trials per subject and the sample size more than the smaller effect sizes (Figs B14).

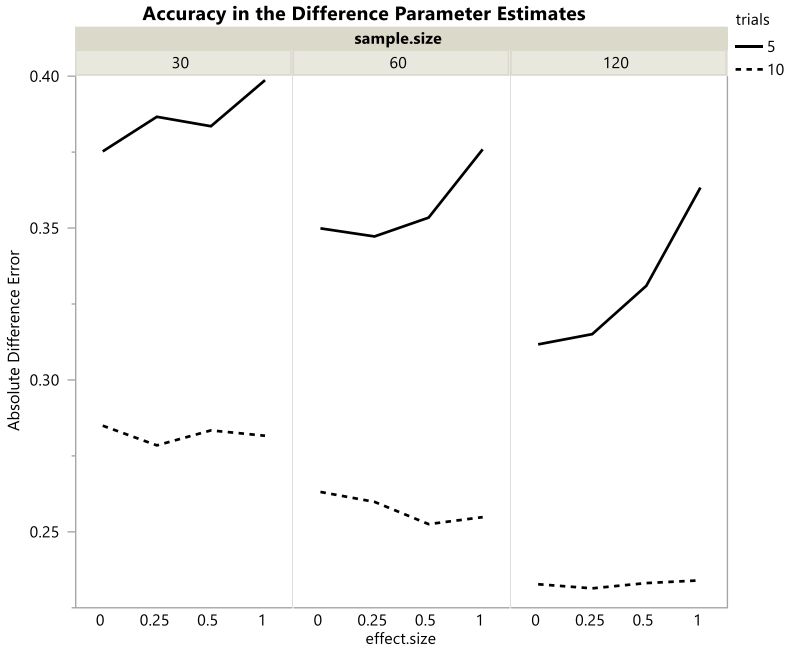

Figure B14 Accuracy in the difference parameter estimates. This figure shows the interaction plot for effect size, sample size and the number of trials per subject for continuous predictors when data are transformed. Each error bar is constructed using 1 standard error from the mean.

In both cases, large effect sizes were consistently less accurate. There was also a three-way interaction when continuous predictors were used but again, this was probably significant because of the tiny mean square error. When discrete predictors were used, the impact of effect size differed depending on which sample size was used.

Applying Gamma Distribution.

Applying a Gamma distribution led to bias in the intercept parameter estimates (Fig B15).

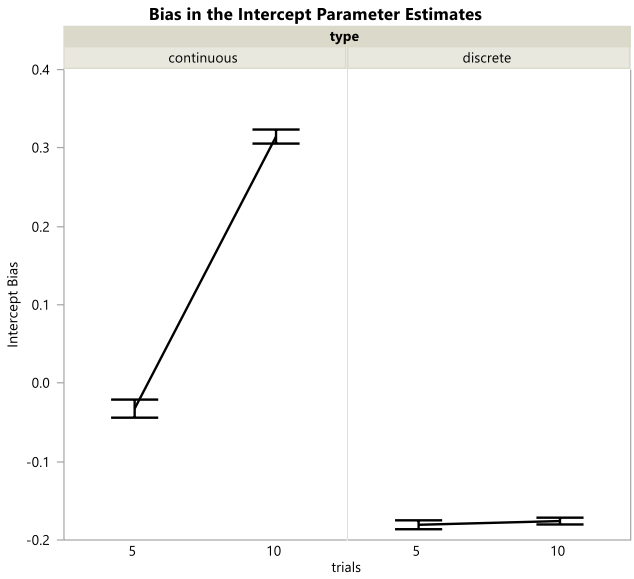

Figure B15 Bias in the intercept parameter estimates. This graph shows the intercept parameter bias by the number of trials per subject when a Gamma distribution is applied. Each error bar is constructed using 1 standard error from the mean.

When continuous predictors were used, the estimates were overestimated, and when discrete predictors were used, the estimates were underestimated. However, these effects were probably significant because of the small standard errors (*SE≈0.006)* and are basically unbiased. There were no other significant effects when discrete predictors were used, but there were several significant effects when continuous predictors were used. Increasing the effect size, $F(3,3)=885.85$, $p<0.001$ and number of trials per subject, $F(1,1)=635.24$, $p<0.0001$, increased the estimation bias for continuous predictors, but neither variable affected the intercept parameter estimates for discrete predictors.

The number of trials per subject also significantly interacted with effect size, $F(3,3)=126.32$, $p<0.001$, for continuous predictors. When increasing the number of trials per subject, each effect size had significantly different impacts on the intercept parameter estimates (Fig B16).

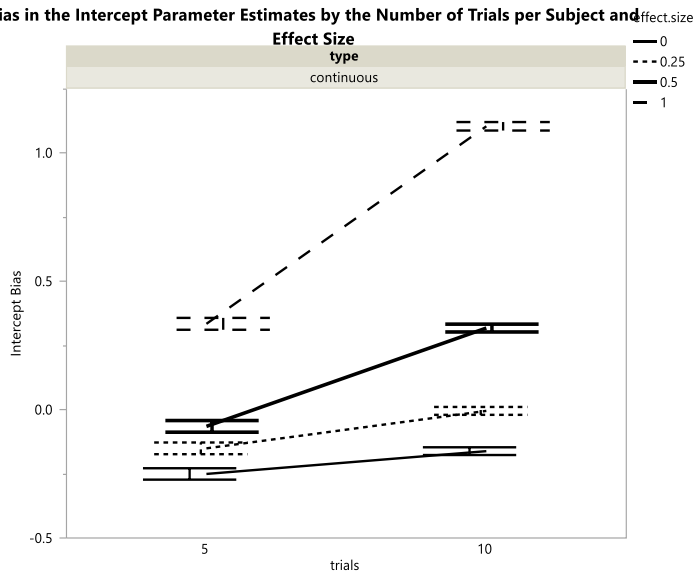

Figure B16 Bias in the intercept parameter estimates by the number of trials per subject and effect size. This figure shows the interaction plot for the number of trials per subject and effect size for bias in the intercept parameter estimates when a Gamma distribution is applied with continuous predictors. Each error bar is constructed using 1 standard error from the mean.

All main effects, for both types of predictors, significantly impacted the accuracy of the intercept parameter estimates. Increasing the number of observations increased accuracy for both types of predictors. Smaller effect sizes were also more accurate than larger ones. There was a significant interaction between the number of trials per subject and sample size for both continuous, $F(2,2)=33.77$, $p<0.0001$, and discrete, $F(2,2)=6.38$, $p=0.0017$, predictors. However, these interactions were negligible when continuous predictors were used and was likely caused by the small mean square error ($MSE=0.009$). For discrete predictors, the smallest sample size improved more quickly than the other sample sizes as trial size increased.

Estimating the difference parameter did not follow the same patterns as the intercept parameter. Applying a Gamma distribution did not bias the difference parameter estimate for discrete predictors. There was significant overestimation when continuous predictors were used, $t=41.42$, $p<0.0001$, but this was surely due to the very small standard error ($SE=0.0002$).

Basically, the difference parameter estimates are unbiased when the Gamma distribution is applied. This technique did result in significant error in the estimates. All main effects and two-way interactions were significant for both types of predictors. The three-way interaction was significant for continuous predictors, but is generally uninterpretable and likely due to the tiny mean square error (*MSE=0.005*). In fact, all the effects are minimal for continuous predictors. Increasing the effect size increased the difference between the number of trials per subject, while increasing sample size did not did not have differential effects on the number of trials per subject (Fig B17).

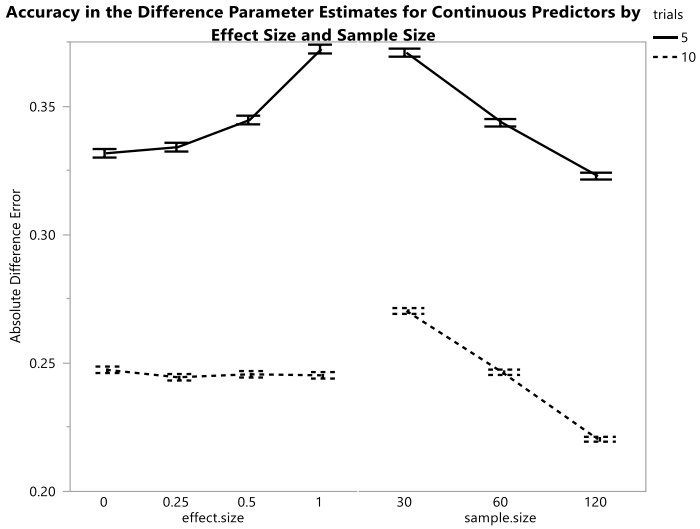

Figure B17 Accuracy in the difference parameter estimates for continuous predictors by effect size and sample size. This figure shows the interaction plot for the number of trials per subject by effect size and sample size when a Gamma distribution is applied. This is for continuous predictors only because there was no significant bias in the difference parameter estimates when discrete predictors were used. Each error bar is constructed using 1 standard error from the mean.

One should take caution, however, in interpreting these two-way interactions as they are most likely caused by very small mean square errors.