Judgment and Decision Making, Vol. 15, No. 5, September 2020, pp. 851–860

The individual true and error model: Getting the most out of limited data

Pele Schramm*

Abstract

True and Error Theory (TET) is a modern latent variable modeling approach for analyzing sets of preferences held by people. Individual True and Error Theory (iTET) allows researchers to estimate the proportion of the time an individual truly holds a particular underlying set of preferences without assuming complete response independence in a repeated measures experimental design. iTET is thus suitable for investigating research questions such as whether an individual ever is truly intransitive in their preferences (i.e., they prefer a to b, b to c, and c to a). While current iTET analysis methods provide the means of investigating such questions they require a lot of data to achieve satisfactory power for hypothesis tests of interest. This paper overviews the performance and shortcomings of the current analysis methods in efficiently using data, while providing new analysis methods that offer substantial gains in power and efficiency.

Keywords: hierarchical, bayesian, true and error, transitivity

1 Introduction

It is of interest to many behavioral decision researchers to determine sets of preferences held by individuals. Indeed, there are many theories that provide for specific constraints on possible sets of preferences one may hold, such as expected utility theory (Allais, 1953), lexicographic semi-orders of preference (Luce, 1956), gain-loss separability (Wu & Markle, 2008), and cumulative prospect theory (Tversky & Kahneman, 1992). Perhaps the most well known constraint is transitivity: for any three options a, b, and c, if a is preferred over b and b is preferred over c, then c cannot be preferred over a. To test such theories, a common experimental approach is to ask people to make repeated binary choices, and then analyze the frequencies of various responses. The majority of such analysis approaches assume that responses across the repeated measures are independent of one another for the sake of statistical convenience (e.g., Tversky (1969), Hey (1995), Regenwetter, Dana, and Davis-Stober (2011)). Birnbaum (2012) demonstrated that this independence assumption can be tested and has been determined to be faulty in some instances. While co-occurrences of preferences are of particular interest when investigating theories such as transitivity, most existing analysis approaches ignore co-occurence of choices. Instead, the tendency is to limit analysis approaches to marginal choice probabilities, i.e., the probabilities of responses to individual binary choices. As pointed out by Birnbaum (2011), these assumptions can lead to incorrect conclusions about people's true underlying sets of binary preferences. If we take, for example, transitivity of preferences, it is possible that people at any given point in time follow transitivity of preference perfectly yet have marginal choice probabilities that reflect a violation of weak stochastic transitivity (that is, if $P(a \succ b) > .5$ and $P(b \succ c) > .5$, then $P(a \succ c) > .5$) if their set of preferences varies at different points in the experiment (Regenwetter et al., 2011). Regenwetter et al. (2011) propose the solution to use the triangle inequality instead of weak stochastic transitivity ($0 \leq P(a \succ b) + P(b \succ c) - P(a \succ c) \leq 1$, but Birnbaum (2011) pointed out that it is possible for people to have a mixture of only intransitive preference orderings without violating the triangle inequality (e.g., if 66% of the time $a \succ b, b \succ c$, and $c \succ a$ and 34% of the time $b \succ a, c \succ b$, and $a \succ c$).

The concern with shortcomings of methods that analyze data as independent binary choices has motivated the development of true-and-error (TE) models, which account for co-occurrence of preferences. TE models originally evolved out of the approach in Lichtenstein and Slovic (1971). The underlying assumption is that, at any given time, an individual has a latent true set of binary preferences, but may respond in a manner inconsistent with their current true set of preferences with a separate error probability possibly ranging from 0 to 0.5 for each binary choice (Birnbaum, 2013). Besides parameters describing the error probabilities for each binary choice, the model includes parameters denoting the probability of a participant holding each possible true set of preferences. In practice, participants are prompted with the same or similar binary choice questions twice in each block

Special thanks to Michael Birnbaum and Jeffrey Rouder for their discussion and input.

*University of California, Irvine, and Technion, Israel Institute of Technology. Cooper Building, Haifa, Israel. E-mail: peleschramm@campus.technion.ac.il.

(e.g., one question might have the options in reverse order), usually with filler questions in between. The model is constrained by the assumption that latent sets of true preferences remain constant within each block, but may vary between blocks. This approach can be used to analyze group data, where each participant completes one block (e.g., Birnbaum, 2007; Birnbaum & Gutierrez, 2007; Birnbaum & Schmidt, 2010), or individual data, where each participant completes multiple blocks (e.g., Birnbaum & Bahra, 2012).

One of the practical limitations of the TE model, especially when applied to individuals separately, is that accurate analysis requires large sample sizes. It's not uncommon for TE experiments to involve multiple sessions each lasting an hour or more in order to achieve the statistical power (rate of correct rejections of a null hypothesis) necessary to reject a set of constraints. Thus, it is of particular interest to researchers using these methods to make the most of the inherently limited amount of data that is available to them. I describe later how the present frequentist approach that is suggested in Birnbaum (2013) and Birnbaum and Quispe-Torreblanca (2018) is suboptimal for efficiently detecting violations of constraints, and will resolve one of the major concerns from a frequentist perspective. After that, both hierarchical and non-hierarchical Bayesian methods of analysis will be explored and consequentially advocated for.

While TE models can be applied to test theories in a number of different domains, such as testing expected utility with the Allais Paradox (Lee, 2018), the focus of this paper is on patterns of preferences among three items, especially dealing with testing whether individuals have truly intransitive sets of preferences. The approaches highlighted in this paper can nonetheless easily be extended to other uses of TE models. Other uses of TE models thus far have included testing dimension integration (Birnbaum & LaCroix, 2008), gain-loss separability (Birnbaum & Bahra, 2007), cumulative prospect theory and the priority heuristic (Birnbaum, 2008), and the Allais Paradox (Birnbaum, 2007).

2 The True-and-Error Model

To understand the important points about the statistical analysis we first need to overview the nature of the data the TE model analyses. When using a TE model to test transitivity, subjects are prompted with the three possible pairwise comparisons between the three items twice per block (i.e., choosing between a and b, b and c, and c and a) for multiple blocks. Thus, there are a total of 64 possible outcomes for each block ($2^3 = 8$ possible sets of preferences for the first iteration of questions times $2^3 = 8$ for the second). For example, in one block a subject might respond 011 for the first set of questions and 001 for the second set of questions, each digit representing a single binary response to the corresponding paired comparison. Matrix A below denotes in each row a separate possible set of preferences. We can think of the first column of A as corresponding to the comparison between items a and b, the second column as b vs. c, and the third as a vs c. In this case, a 0 can represent having chosen the item earlier in the alphabet and a 1 the item later in the alphabet, so that rows 2 and 7 represent intransitive preference sets.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{1}$$

Now we can define $p_{i,j}$ to be the probability of subject i holding the true set of preferences corresponding to the jth row in A in any one block. These are usually the primary parameters of interest. Formally, if we let $T_{i,m}$ denote the index of the row of A corresponding to the true set of preferences subject i holds in block m, then:

$$P(T_{i,m} = j) = p_{i,j} \tag{2}$$

.

If we let $f_{i,m}$ and $g_{i,m}$ denote the index of the row of A corresponding to the observed set of preferences subject i reports in the first and second set of questions in block m, and let $e_{i,k}$ denote the probability of error in reporting the true latent set of preferences for subject i for paired comparison k, we have:

$$\begin{aligned} P(f_{i,m}|T_{i,m}) = \prod_{k=1}^{3} I(A_{f_{i,m},k} = A_{T_{i,m},k})[1 - e_{i,k}] \\ + [1 - I(A_{f_{i,m},k} = A_{T_{i,m},k})]e_{i,k}, \end{aligned} \tag{3}$$

and

$$\begin{aligned} P(g_{i,m}|T_{i,m}) = \prod_{k=1}^{3} I(A_{g_{i,m},k} = A_{T_{i,m},k})[1 - e_{i,k}] \\ + [1 - I(A_{g_{i,m},k} = A_{T_{i,m},k})]e_{i,k}, \end{aligned} \tag{4}$$

where I denotes an indicator taking values of 0 if the statement inside is incorrect, and 1 if correct. The expression within the product over k is equivalent to the probability of having observed the k^{th} paired comparison given a specified true preference pattern according to the model, which is simply $1 - e_{i,k}$ when it corresponds to the true preference pattern and $e_{i,k}$ when it doesn't.

Because the model assumes that $T_{i,m}$ remains constant for all six preference indications in $f_{i,m}$ and $g_{i,m}$, Ts can be marginalized out completely from the model using the law of total probability. Thus we can treat the combination of the two sets of preferences observed in a block as having a joint probability following:

$$P(f_{i,m}, g_{i,m}) = \sum_{j=1}^{8} p_{i,j} P(f_{i,m}|T_{i,m} = j) P(g_{i,m}|T_{i,m} = j). \quad (5)$$

Perhaps it should be stressed that this model lacks any underlying concept of utility. Instead, the model deals purely with probabilities of concurrent sets of preferences and error probabilities for each pairwise comparison. From these parameters it is possible to calculate the marginal probabilities of each pairwise comparison, but the model's strengths lie in its ability to look beyond these concepts and conduct analysis without the restriction of analyzing only marginal paired comparison probabilites. Consequentially there is no direct mapping between the TE model and random utility models such as the Thurstonian (or probit) model (Thurstone, 1927) or the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959). What might be represented as two items being close in utility under a random utility model can be represented by either competition between preference sets where the relative ranking of the two items differ, or by a high error term e for that paired comparison. This is determined according to the frequencies of observed preference reversals within and between blocks, which is something that random utility models do not consider.

3 Shortcomings and Improvement to the Present Frequentist Approach

For the sake of fitting and testing TE models, Birnbaum (2013) suggests reducing the degrees of freedom in the data down to 15 by only looking at the first set of preferences per block and noting whether the second set matches the first perfectly. For the purpose of hypothesis testing (comparing a less constrained null TE model vs a more constrained TE model), a Pearson's chi-squared test (Pearson, 1900) is proposed by plugging in the chi-squared statistics on the 16 observed frequencies for an unrestricted model vs a restricted model (e.g., one which has a fixed zero probability of true intransitivity). The null distribution of the difference in chi-square statistics in this case is said to come from a chi-square distribution with degrees of freedom equal to the difference in number of estimated TE parameters between the restricted and unrestricted model. While this approach is highlighted, there is acknowlegment that in some cases use of the full data can be more appropriate, and the full data were used in Birnbaum and Quispe-Torreblanca (2018).

There are two major problems with Birnbaum's degrees of freedom reduction approach, one that is easy to resolve and one that is less so. Perhaps the most substantial issue is that in the reduction of the degrees of freedom, a substantial (and indeed, useful) part of the data collected is left unaccounted for (i.e., an entire set of observed preferences gets reduced to whether it was the same or different from the other one). Because data are practically limited, especially in the case of analyses on the individual level, this turns out to be a great sacrifice.

The motivation behind the reduction in the degrees of freedom seems to be to make the chi-square test a feasible option. Unfortunately, the specified Pearson's chi-square test is not appropriate for these purposes, and turns out to be overly conservative as demonstrated by the simulations in the following section. The result is that the true type I error rate (rate of false rejections) is far lower than the nominal α level, and p-values appear higher than they should be. Besides potential issues of small sample sizes for frequency data with 15 degrees of freedom, the Pearson chi-square test is supposed to feature a chi-square null distribution with degrees of freedom equal to the difference in number of outcome probabilities fixed. In the case of the TE model, specific outcome probabilities aren't being fixed. Instead model parameters that have some effect on potentially all outcome probabilities are being fixed. Thus, it turns out that this reduction in degrees of freedom comes at great cost while not fulfilling the original purpose.

Luckily, the likelihood of the full, unreduced data is easy to calculate by multiplying all the probabilities of each observed block shown in equation 5, and so a potential alternative would be a Likelihood Ratio Test (Neyman & Pearson, 1933). The famous Neyman Pearson Lemma introduced in Neyman and Pearson (1933) proves that the Likelihood Ratio Test is the single most powerful test when comparing nested models. While the exact distribution of the test statistic (twice the log of the likelihood ratio) is often difficult to derive, Wilks (1938) showed that, just like the Pearson chi-squared test, the Likelihood Ratio test statistic is asymptotically distributed according to a chi-square distribution under the null under certain regularity conditions. Unfortunately this isn't guaranteed when parameters are being fixed at the end points of their possible ranges, and null TE models usually feature p parameters fixed at 0, their minimum possible value. Even when parameters are not being fixed at 0 or 1, it may take many blocks to reach the asymptotic limit, more than one could hope to get from one individual. Although this approach has the advantage of making use of all of the data, it fails to resolve the issue with the chi-square test of a questionable type I error rate and thus inaccurate p-values. Despite the theoretical issues surrounding these applications of the likelihood ratio and chi-square tests, rejections of the null hypothesis from either test appear trustworthy, as will be demonstrated via the simulations in the following section.

On the contrary, these tests were both found to be overly conservative, so elevated Type II errors (i.e., failures to detect violations of transitivity) are the primary concern.

To avoid direct reliance on theoretical test distributions, Birnbaum et al. (2016) implemented a bootstrapping procedure to calculate confidence intervals of parameter estimates, and a Monte-Carlo simulation procedure for estimating the distribution of test statistics. Bootstrapping is perfomed by iteratively refitting the model with many datasets sampled from the original dataset with replacement to yield a distribution of parameter estimates. Monte Carlo simulation is performed by fitting the model and then generating simulated datasets from the parameter estimates. Since we are interested in the null distribution of the test statistics, a Monte-Carlo approach to hypothesis testing could be to fit the null model and generate samples from the parameter estimates, looking at the distribution of test statistics and checking whether the raw test statistic falls outside this range. While Birnbaum et al. (2016) used the reduced data approach to fit the bootstrapped and Monte Carlo simulated data sets, I investigate in this paper the efficacy of using the full data approach with Likelihood Ratio Test statistics.

3.1 Simulations

To explore power, type I error rate, and accuracy of parameter estimation, two separate simulation strategies were employed to generate the parameters representing the probabilities of a subject holding each possible set of true preferences in a block. The first one, which will be referred to as the probit simulation, uses a probit model, or Thurstonian Case V (Thurstone, 1927) to generate the probabilities of the true sets of preferences (the p parameters of the model). The three items, which we can call a, b, and c, were given average probit values of -1, 0, and 1 respectively. For each simulated participant, their personal probit values were drawn from standard normal distributions centered at these 3 values. The probability of each true set of preferences was assigned according to the corresponding probability of observing that set of preferences if they were responding in accordance with a probit model. For example, the probability for the true set of preferences being a is preferred to b, b is preferred to c, and a is preferred to c would be $\Phi(V_a - V_b) \times \Phi(V_b - V_c) \times \Phi(V_a - V_c)$ where V_x is the probit value for item x for that individual and Φ is the cumulative distribution function of the standard normal. The three error probabilities were each set to 1/2 times a value independently drawn from a $Beta(1, 2)$ distribution, slightly favoring lower error rates. It should be noted that while the probit model is used, the actual marginal probabilities of responses to single prompts does not reflect the probit model, but the true and error model. The decision to use a probit model was purely in hopes of generating sets of TE parameters which resemble some underlying random utility structure.

The second method of simulation was more flexible and general. Later in the paper a hierarchical model is going to be introduced that exploits an assumption of similarity among people's parameter values to gain better estimates, so simulation in this case is done on a group level. For the purposes of the frequentist tests, the group size is simply set to 1 since there is no built-in hierarchical structure in the model anyway. Initially, a single vector was drawn from an 8 outcome flat Dirichlet distribution, which can be thought of as an 8 dimensional extension to the uniform distribution where all elements sum to 1. After that, each subject's true probabilities' values were drawn from a Dirichlet parameterized by that initial vector multiplied by a single random variable distributed as a $Gamma(8, 1)$, which is a continuous distribution defined from 0 to ∞ with a mean of 8 and a variance of 8. This gamma random variable represents the concentration parameter of the Dirichlet distribution. The concentration parameter dictates the expected sparcity of drawn values, with larger values indicating a bias toward resulting vectors that are more even and smaller values indicating a bias toward vectors concentrated on a small proportion of the elements. When the concentration parameter is equal to the dimensionality of the Dirichlet, in this case 8, there is no bias of this nature. For our purposes, a concentration parameter that is much higher than 8 can be understood as representing an expectation that the probabilities of all 8 preference sets will be approximately equal to one another, while a concentration parameter that is much less than 8 represents an expectation that only one or two preference sets will be dominant. The three error probabilities were simulated in the same way as before.

For simulations geared toward detection of transitivity or intransitivity, transitive individuals had the two p parameters corresponding to intransitive sets of preferences (those corresponding to the 2nd and 7th rows of matrix A in equation 1) set to 0 following the afforementioned generation strategies, and then their p vector was renormalized. For the case of intransitivity, values generated from a Normal distribution centered at 1 with standard deviation of 0.2 were added to the p parameter corresponding to $b \succ a, c \succ b, a \succ c$ and then the entire vector was renormalized. To help conceptualize this, if one starts out with a probability of 0 of having the aforementioned intransitive pattern and a 1 gets added to it, the probability of holding that true preference pattern becomes 0.5 after renormalization.

Data with 12 blocks per person and with 24 blocks per person were simulated in each of the cases, 12 representing a relatively small number of data one would collect with the individual model and 24 representing a relatively large number of data. For simulations involving the hierarchical model defined later, each case included simulations with 15 simulated participants and 60 simulated participants to illustrate differences in performance when data is available from more subjects.

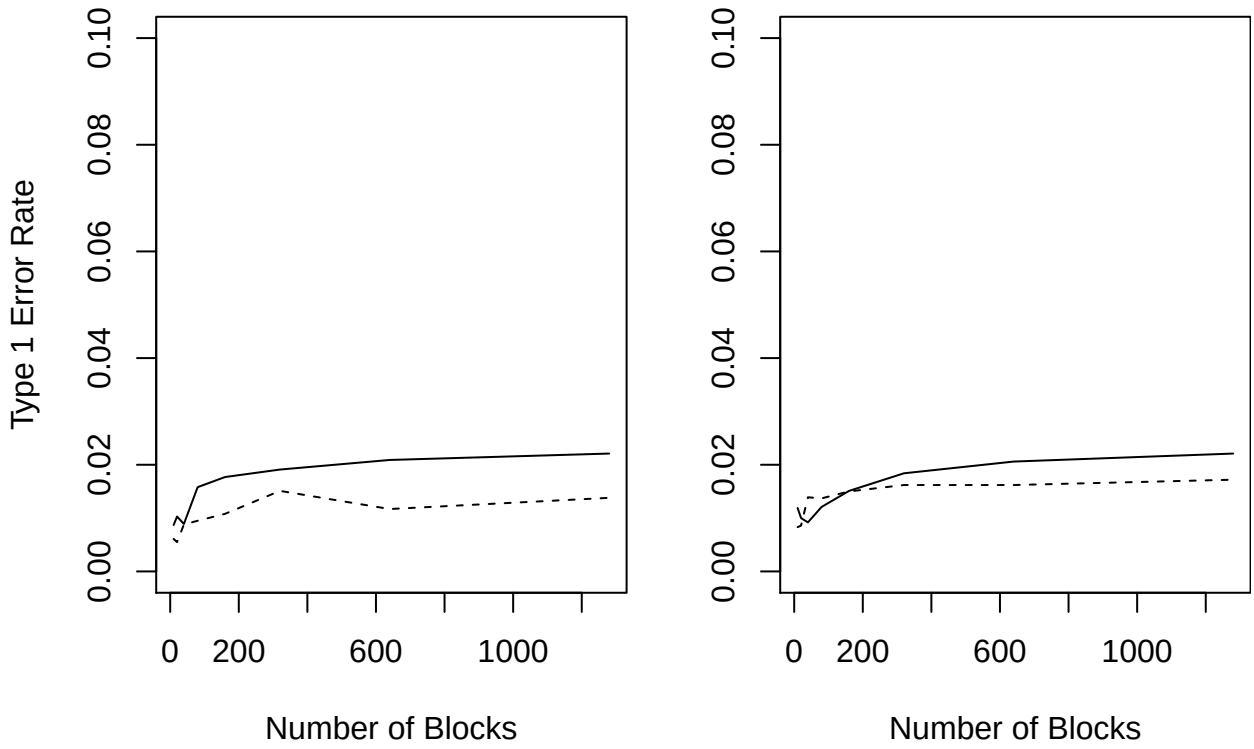

FIGURE 1: Type 1 Error Rate vs. Number of Blocks for the Likelihood Ratio Test (dashed) and Chi Square (solid). On the left are results via simulation with the probit parameter generation approach, and on the right the Dirichlet approach.

The same simulated datasets were used for power analyses of bootstrapping and Monte-Carlo procedures. Each case of bootstrapping used 1000 bootstrapped datasets, and Monte-carlo also used 1000 Monte-Carlo samples. The null hypothesis was said to be rejected in the case of bootstrapping if the 95% confidence interval of either of the two possible intransitive probabilities did not include any value below .001. Although it is not recommended to use confidence intervals for testing point null hypotheses (Wagenmakers, Lee, Rouder & Morey, 2019), this was done to correspond with the procedure used in Birnbaum et al. (2016). For the Monte-Carlo procedure, the null was said to be rejected if 95% or more of the simulated LRT statistic distribution was below the real LRT statistic. 2000 experiments were simulated for each condition using the University of California, Irvine's high performance computing cluster.

3.2 Results

The Mean Squared Error (MSE) of the estimates of the resulting true preference set probabilities is given in the rightmost two columns in Table 1 both using the full data (on the left) and data reduced as suggested in Birnbaum (2013). As we can clearly see, making use of the full data provides a substantial reduction in MSE, yielding more accurate estimates with the same number of blocks.

To check whether the true type I error rates for a nominal $\alpha = .05$ converge to the nominal value for the Likelihood Ratio Test and chi-squared tests, type I error was estimated via simulation for different numbers of blocks. These results are shown in Figure 1. These simulations show that both tests stay far below their nominal α value for any realistic block size for individuals. The chi-squared test type 1 error seems to increase somewhat faster than the likelihood ratio test but even past 1000 blocks neither of them seem to level off at the nominal α level.

A power analysis for the chi-square, LRT, bootstrap, and Monte Carlo approaches can be found in Table 2. We clearly see that despite the true type I error rate being slightly worse for the LRT, the gains in power are substantial relative to chi-square. While we don't really know the true null distributions for these tests, it seems like we can trust rejections. While still being overly conservative, the bootstrap procedure yields by far the highest power. Although the Monte-Carlo method has a type I error rate that is close to the nominal alpha level, its power is worse than the other methods. This is not too surprising, because in a sense the Monte-Carlo simulations use what could be thought of as a worst-case scenario null distribution, selecting the most likely set of null parameters to have generated the original data.

TABLE 1: Mean Squared Error of probability estimates for each estimation method. For the Bayesian results, MSE(est) denotes the MSE with respect to the posterior mean, while MSE(post) denotes the MSE with respect to the posterior distribution. MSE(full) denotes the MSE with respect to a maximum likelihood fit using all the data, while MSE(red) denotes the MSE with respect to a fit using reduced data as in Birnbaum (2013).

		Hierarchical Bayes		Individual Bayes			
N Subjects	N Blocks	MSE(est)	MSE(post)	MSE(est)	MSE(post)	MSE(full)	MSE(red)
Probit							
15	12	0.0114	0.0232	0.0202	0.0276	0.0173	0.0242
15	24	0.0069	0.0147	0.0126	0.0180	0.0107	0.0150
60	12	0.0100	0.0208	0.0205	0.0279	0.0174	0.0238
60	24	0.0069	0.0144	0.0128	0.0182	0.0113	0.0155
Dirichlet							
15	12	0.0085	0.0175	0.0124	0.0201	0.0186	0.0244
15	24	0.0059	0.0123	0.0089	0.0147	0.0115	0.0139
60	12	0.0078	0.0154	0.0125	0.0202	0.0183	0.0230
60	24	0.0055	0.0109	0.0091	0.0149	0.0113	0.0144
Hier							
15	12	0.0096	0.0197	0.0186	0.0260	0.0171	0.0218
15	24	0.0061	0.0123	0.0129	0.0183	0.0105	0.0139
60	12	0.0088	0.0176	0.0203	0.0277	0.0179	0.0229
60	24	0.0059	0.0116	0.0126	0.0180	0.0109	0.0138

TABLE 2: Power and level for each frequentist hypothesis testing method for both parameter generation approaches. Power is the proportion of correct rejections of the transitive null model for intransitive simulations, and level is the proportion of false rejections of transitivity for transitive simulations, each at the nominal $\alpha = .05$.

	LRT		Chi Square		Bootstrap		Monte Carlo	
N Blocks	Power	Level	Power	Level	Power	Level	Power	Level
Probit								
12	0.396	0.008	0.292	0.009	0.439	0.004	0.288	0.040
24	0.564	0.007	0.444	0.013	0.698	0.003	0.382	0.049
Dirichlet								
12	0.447	0.009	0.372	0.011	0.513	0.005	0.375	0.040
24	0.624	0.010	0.497	0.012	0.756	0.006	0.460	0.042

4 Bayesian Hierarchical Model

Bayesian hierarchical models in a sense allow behavioral researchers to get the best of both worlds: analysis on an individual subject level while still making use of group-level information. Hierarchical models are powerful tools that have been proven to provide more accurate parameter estimates than non-hierarchical models, as measured by MSE (Efron & Morris, 1977) by formalizing stronger distributional assumptions. For example, a hierarchical model might specify a probability distribution for person-specific parameters, whereas a non-hierarchical model would estimate person-specific parameters independently for each person, without considering that there may be similarities across people. Bayesian statistics in general differs from frequentist statistics in that, rather than providing point estimates for parameters, posterior distributions are inferred from the data according to a model specification which incorporates

prior assumptions on the parameters (Gelman et al., 2013; Wagenmakers, Lee, Lodewyckx & Iverson, 2008).

Although it was applied to a different variation of the True and Error model involving only two preferences per set instead of three, Lee (2018) implemented a non-hierarchical Bayesian analysis of the True and Error model. Since our goal is to get the most out of our limited data, a hierarchical model is a natural expansion of this approach. The one used in this paper shares the same cognitive model as defined previously, but with hierarchical priors for individual p parameters. More specifically, a soft-max transformation of normally distributed latent variables is employed:

$$X_{i,j} \sim \text{Normal}(\mu_j, \tau_j) \tag{6}$$

$$p_{i,j} = \frac{e^{X_{i,j}}}{\sum_{j=1}^{8} e^{X_{i,j}}} \tag{7}$$

$$\mu_j \sim \text{Normal}(0, 1) \tag{8}$$

$$\tau_j \sim \text{Exponential}(1) \tag{9}$$

Here, parameters subscripted with j represent belonging to the particular set of preferences in the jth row of A, and i denotes the subject number. The normal distributions are parameterized according to precision (that is, inverse variance). The essence of this hierarchical model is that values ($X_{i,j}$) for each probability of a particular set of preferences are drawn from the same normal distributions across all subjects. These X values can be thought of as normally distributed representations of the probabilities for each subject to hold each possible underlying set of preferences in any given block. The probabilities for subject i to hold each underlying set of preferences in any given block can be calculated directly from these normally distributed representations via softmax as in equation 7.

The error probabilities, denoted by e, are all halves of Beta(1,2) distributed random variables since it makes sense to assume low errors are more probable than ones nearing chance:

$$2e_{i,k} \sim Beta(1, 2) \tag{10}$$

This specification is a slight departure from Lee (2018)'s implementation, where the error parameters were drawn from uniform distributions.

The non-hierarchical model used in this paper differs from the hierarchical model only in that the ps are instead drawn from a flat Dirichlet distribution, which can be thought of as a multidimensional uniform distribution. The following specification for the non-hirarchical model replaces equations 6 through 9:

$$p_{i,1:8} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1) \tag{11}$$

An implementation of these models in JAGS is given in the appendix.

MSE performance for estimating the ps in each model from 100 simulations of each case is shown in Table 1, both in posterior distribution as well as in point estimation (in this case the mean of the posterior distribution). The simulation approaches previously highlighted do not correspond perfectly to the hierarchical specification. The mismatch between hierarchical specification and simulations was done deliberately to see whether the hierarchical model was robust to misspecification of hierarchical structure. For the sake of comparison, simulations that do correspond precisely to the hierarchical model, drawn directly from the specified priors, were also performed. The results show that the hierarchical Bayesian model tends to outperform the Individual model substantially in all cases, often by a factor of 2. We also see moderate performance gains with the hierarchical model in cases with 12 blocks per person when 60 subjects are included in analysis vs only 15. It's noteworthy that the hierarchical model here outperforms all other models even for the simulation approaches that do not generate people's parameters according to softmax transformed normally distributed variables as specified. Interestingly, it appears as though these mismatches between the true generating process and the specified one made no discernible difference in performance.

4.1 Bayesian Hypothesis Testing

The favored approach by Bayesians for hypothesis testing is the Bayes factor, which can be conceptualized as the ratio of the expectation of the probability of the data over the prior distributions of the two models being compared (Jeffreys, 1961). In this case, one can practically calculate the Bayes factor using a spike-and-slab approach by adding a Bernoulli distributed indicator parameter h for transitivity with prior probability of 1/2. When h is 1, then the ps corresponding to a set of preferences which violate transitivity are automatically set to 0, and otherwise they are said to come from the distribution denoted in the above model specification. For the hierarchical model, the following specification replaces equation 7:

$$h_i \sim \text{Bernoulli}(0.5) \tag{12}$$

$$p_{i,j} = \frac{I[(j \neq 2, 7 \& h_i = 1) or h_i = 0] e^{X_{i,j}}}{\sum_{j=1}^{8} \left(I[(j \neq 2, 7 \& h_i = 1) or h_i = 0] e^{X_{i,j}} \right)} \tag{13}$$

For the non-hierarchical model, this can be implemented by drawing from separate Dirichlet's in each case, flat in the intransitive case, and with zeros for the intransitive parameters and $\frac{4}{3}$ for the others in the transitive case. The following specification replaces equation 11:

$$h_i \sim \text{Bernoulli}(0.5) \tag{14}$$

$$q_{i,1:8} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1, 1, 1) \tag{15}$$

$$u_{i,1:8} \sim \text{Dirichlet}(\tfrac{4}{3}, 0, \tfrac{4}{3}, \tfrac{4}{3}, \tfrac{4}{3}, \tfrac{4}{3}, 0, \tfrac{4}{3}) \tag{16}$$

$$p_{i,1:8} = h_i u_{i,1:8} + (1 - h_i) q_{i,1:8} \tag{17}$$

TABLE 3: Hypothesis test results for the two Bayesian models. "C" denotes the proportion whose Bayes Factors favor the right direction, BF>x denotes the proportion of intransitive people with a Bayes Factor greater than x favoring intransitivity, and BF>xF denotes the proportion who were transitive yet still had a Bayes Factor greater than x favoring intransitivity.

		Hierarchical Bayes					Individual Bayes				
S	B	C	BF>3	BF>3F	BF>10	BF>10F	C	BF>3	BF>3F	BF>10	BF>10F
Probit											
15	12	0.88	0.72	0.03	0.57	0.01	0.87	0.72	0.05	0.49	0.01
15	24	0.96	0.85	0.03	0.70	0.02	0.92	0.88	0.03	0.73	0.02
60	12	0.93	0.83	0.03	0.69	0.00	0.86	0.76	0.04	0.56	0.01
60	24	0.94	0.89	0.02	0.76	0.01	0.90	0.86	0.05	0.70	0.01
Dirich											
15	12	0.91	0.83	0.04	0.71	0.01	0.85	0.80	0.06	0.60	0.00
15	24	0.94	0.95	0.03	0.88	0.01	0.89	0.92	0.04	0.81	0.01
60	12	0.93	0.87	0.03	0.77	0.02	0.86	0.81	0.05	0.61	0.01
60	24	0.96	0.94	0.02	0.89	0.01	0.90	0.92	0.05	0.82	0.02
Hier											
15	12	0.90	0.76	0.03	0.57	0.00	0.84	0.72	0.04	0.53	0.01
15	24	0.93	0.91	0.02	0.84	0.00	0.89	0.92	0.06	0.80	0.01
60	12	0.92	0.85	0.04	0.71	0.01	0.86	0.80	0.06	0.60	0.02
60	24	0.95	0.93	0.02	0.86	0.01	0.89	0.91	0.06	0.78	0.02

The proportion of the time the indicator shows intransitivity ($h = 0$) divided by the proportion of the time the indicator indicates transitivity ($h = 1$) is the Bayes Factor for that individual being intransitive.

To explore relative performance of the hierarchical vs non-hierarchical formulations of the model for Hypothesis testing, simulations were done as before (50 times per each case), this time where each subject had a 0.5 chance of being truly transitive or intransitive. Results of this can be seen in Table 3. While some researchers might want to avoid having formal cutoffs for Bayes Factors (de Vries & Morey, 2013), proportion of Bayes Factors greater than 1, 3 and 10 were reported along with corresponding type 1 error rates. Wagenmakers, Morey and Lee (2016) suggest that Bayes factors greater than 1 correspond to "anecdotal" evidence, 3 tend to correspond to a "moderate" amount of evidence, and greater than 10 a "strong" amount of evidence.

We can see substantially better performance here than in the frequentist tests, even with highly conservative cutoffs at 10, and substantially better performance for the hierarchical implementation relative to the individual Bayesian implementation. From these simulations, we see a Bayes factor of 3 in favor of a transitivity violation approximately corresponds to a type-1 error rate of 0.05 in the case of the non-hierachical individual model, and slightly more conservative than that for the hierarchical model, while a Bayes Factor of 10 has a type-1 error rate of around 0.01. Despite the type 1 error rate of BF>10 being somewhat similar to the frequentist tests from the previous section, we see higher powered results in both Bayesian implementations. Similarly to MSE, the departures from the hierarchical model's specification in the simulated data generating procedure made no discernible difference to performance.

5 Conclusion

In cases where the individual TE model is employed and there are multiple participants responding to the same stimuli, the hierarchical TE model that was introduced in this paper seems to yield the best results. In other cases, there is little reason to use the frequentist approach over the non-hierarchical Bayesian approach, even when properly using all the data, because the proper null distributions are still unknown and their explored frequentist tests tend to be overly conservative. While bootstrapping and using all of the data gives a substantial gain in performance, it still relies on faulty statistical theory for hypothesis testing, and ultimately fails to compete with the Bayesian approach in estimation accuracy and power.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, 503–546.

Birnbaum, M. H. (2007). Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes*, *102*(2), 154–173.

Birnbaum, M. H. (2008). New tests of cumulative prospect theory and the priority heuristic: Probability-outcome tradeoff with branch splitting. *Judgment and Decision Making*, *3*(4), 304–316.

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*(4), 675–683.

Birnbaum, M. (2012). A statistical test of the assumption that repeated choices are independently and identically distributed. *Judgment and Decision Making*, *7*, 97–109.

Birnbaum, M. H. (2013). True-and-error models violate independence and yet they are testable. *Judgment and Decision Making*, *8*(6), 717–737.

Birnbaum, M. H., & Bahra, J. P. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science*, *53*(6), 1016–1028.

Birnbaum, M. H., & Bahra, J. P. (2012). Testing transitivity of preferences in individuals using linked designs. *Judgment and Decision Making, 7*, 524–567.

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organizational Behavior and Human Decision Processes*, *104*(1), 96–112.

Birnbaum, M. H., & LaCroix, A. R. (2008). Dimension integration: Testing models without trade-offs. *Organizational Behavior and Human Decision Processes*, *105*(1), 122–133.

Birnbaum, M. H., Navarro-Martinez, D., Ungemach, C., Stewart, N., Quispe-Torreblanca, E. G.. (2016). Risky decision making: Testing for violations of transitivity predicted by an editing mechanism. *Judgment and Decision Making*, *11*(1), 75–91.

Birnbaum, M. H., & Quispe-Torreblanca, E. G. (2018). TEMAP2.R: True and error model analysis program in R. *Judgment and Decision Making*, *13*(5), 428–440.

Birnbaum, M. H., & Schmidt, U. (2010). Testing transitivity in choice under risk. *Theory and Decision*, *69*(4), 599–614.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods*, *18*(2), 165–185.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Hey, J. D. (1995). Experimental investigations of errors in decision making under risk. *European Economic Review*, *39*(3–4), 633–640.

Jeffreys, H. (1961). *Theory of probability*. Clarendon, Oxford.

Lee, M. D. (2018). Bayesian methods for analyzing true-and-error models. *Judgment and Decision Making*, *13*(6), 622–635.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46–55.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica, Journal of the Econometric Society*, 178–191.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42–56.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer.

Wagenmakers, E.-J., Lee, M., Rouder, J. N., & Morey, R. D. (2019). The principle of predictive irrelevance, or why intervals should not be used for model comparison featuring a point null hypothesis. https://psyarxiv.com/rqnu5.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current*

Directions in Psychological Science, *25*(3), 169–176.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Wu, G., & Markle, A. B. (2008). An empirical test of gain-loss separability in prospect theory. *Management Science*, *54*(7), 1322–1335.

Appendix: JAGS Code:

The following JAGS code was used to analyze the hierarchical model. For the non-hierarchical version, one can remove the lines labeled "Hierarchical Only" and uncomment the line labeled "Non-Hierarchical Only".

Response data is represented by a nobs×2 matrix "fg" where nobs is the number of total blocks. Each row of fg has two integers between 1 and 8 representing the observed preference patterns (defined by the rows of matrix A) for the first and second repetitions in the block. Subject id is represented by the nobs ×1 vector s where each element denotes the subject id of the corresponding block. For the sake of code brevity, this implementation also requires uploading a nobs×1 vector of ones called "onevec", which in the JAGS code is said to come from a Bernoulli distribution with probability equal to the probability of having observed the outcome of that block. Coding it this way is equivalent to calculating the probability of all 64 possible combinations of preference orderings beforehand and treating the combination as coming from a categorical distribution, similar to the approach found in Lee (2018).

```
model{

  for(i in 1:nsub){
    for(j in 1:8){
      X[i,j] ~ dnorm(mu[j], tau[j]) #Hierarchical Only
      expX[i,j] <- exp(X[i,j])      #Hierarchical Only
    }
    ps[i,1:8]<- expX[i,1:8]/sum(expX[i,1:8]) # Hierarchical Only
    #ps[i,1:8]<- ddirch(c(1,1,1,1,1,1,1,1)) #Non-Hierarchical Only
  }

  for(i in 1:8){
    mu[i] ~ dnorm(0,1)
    tau[i] ~ dgamma(1,1)
  }


  for(su in 1:nsub){
    for(i in 1:3){
      doubes[su,i] ~ dbeta(1,2)
      es[su, i] <- .5*doubes[su, i]
    }
  }

  A[1, 1:3] <-c(0,0,0)
  A[2, 1:3] <-c(0,0,1)
  A[3, 1:3] <-c(0,1,0)
  A[4, 1:3] <-c(0,1,1)
  A[5, 1:3] <-c(1,0,0)
  A[6, 1:3] <-c(1,0,1)
  A[7, 1:3] <-c(1,1,0)
  A[8, 1:3] <-c(1,1,1)

  for(h in 1:nobs){

    for(i in 1:8){
      probcomp[h,i] <- ps[s[h],i]*(ifelse(A[i,1]==A[fg[h,1],1], 1-es[s[h], 1],
      es[s[h],1])*ifelse(A[i,2]==A[fg[h,1],2], 1-es[s[h],2], es[s[h], 2])
      *ifelse(A[i,3]==A[fg[h,1],3], 1-es[s[h],3], es[s[h], 3]))
      *(ifelse(A[i,1]==A[fg[h,2],1], 1-es[s[h], 1], es[s[h],1])
      *ifelse(A[i,2]==A[fg[h,2],2], 1-es[s[h],2], es[s[h], 2])
      *ifelse(A[i,3]==A[fg[h,2],3], 1-es[s[h],3], es[s[h], 3]))
    }

    onevec[h] ~ dbern(sum(probcomp[h,1:8]))
  }

}
```