UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Analysis of Cognition at the Individual Level

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Science

by

Pele Schramm

Dissertation Committee:
Jeffrey N. Rouder, Chair
Mark Steyvers
Zygmunt Pizlo

2019

DEDICATION

Dedicated to the memories of Oded Schramm and Bill Batchelder

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thanks to:
Jeffrey Rouder
Bill Batchelder
Michael Birnbaum
Michael Lee
Mark Steyvers
Zygmunt Pizlo
Tselil Schramm
Avivit Schramm

Chapter 1 was done in collaboration with Jeffrey N. Rouder, Chapter 2 with William H. Batchelder, and Chapter 4 with Michael H. Birnbaum.

Supported by NSF Grant #1534471

CURRICULUM VITAE

Pele Schramm

EDUCATION

Doctor of Philosophy in Cognitive Science — **2019**
University of California, Irvine — *Irvine, CA*

Master of Science in Statistics — **2019**
University of California, Irvine — *Irvine, CA*

Bachelor of Science in Physics — **2014**
Western Washington University — *Bellingham, WA*

Bachelor of Arts in Psychology — **2014**
Western Washington University — *Bellingham, WA*

RESEARCH EXPERIENCE

Graduate Student Researcher — **2014–2019**
University of California, Irvine — *Irvine, California*

Undergraduate Student Research assistant — **2012–2014**
Western Washington University — *Bellingham, WA*

Undergraduate Student Research assistant — **2013**
University of Washington — *Seattle, WA*

TEACHING EXPERIENCE

Teaching Assistant — **2014-2019**
University Of California, Irvine — *Irvine, CA*

Physics Lab Teaching Assistant — **2012-2014**
Western Washington University — *Bellingham, WA*

CONFERENCE PRESENTATIONS

Measuring Subjective Value Functions Across Quantitative Multi-Attribute Spaces — **Aug 2016**
Society of Mathematical Psychology Meeting

Cultural Consensus Theory Models for Paired-Comparisons (presented by William H. Batchelder) — **Aug 2016**
Society of Mathematical Psychology Meeting

Subjective Value Function Surface Circus — **Feb 2017**
Edwards Bayesian Research Conference

A Thurstonian Approach to Probabilistic and Temporal Discounting — **Jul 2017**
Society of Mathematical Psychology Meeting

Curvature Agnostic Measurement of Multi-Attribute Subjective Value Functions — **Aug 2017**
Subjective Probability, Utility, and Decision Making

Is a Dime Worth the Time? — **Feb 2018**
Society of Mathematical Psychology Meeting

Investigation into the Transitivity of Preference in Probabilistic and Temporal Discounting — **Jul 2018**
Society of Mathematical Psychology Meeting

Hierarchical Paired Comparison Modeling, A Cultural Consensus Theory Approach — **Jul 2019**
Society of Mathematical Psychology Meeting

ABSTRACT OF THE DISSERTATION

Statistical Analysis of Cognition at the Individual Level

By

Pele Schramm

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2019

Jeffrey N. Rouder, Chair

The focus of this dissertation pertains to effective statistical analysis of cognition at the individual level. The first chapter challenges conventional wisdom for response time analysis by investigating whether there is any benefit to log or reciprocal transforming response times when doing a conventional t-test. The second chapter introduces two models for paired comparison data based on Cultural Consensus Theory. Through hierarchical Bayesian modeling, these models allow for recovery of parameters describing both group level and individual level opinion, tendency toward agreement, and consistency of evaluation as it pertains to the items being compared. The third chapter offers critique and improvements to individual-level True and Error analysis, a modern statistical framework for the evaluation of concurrent sets of preferences. A Hierarchical Bayesian implementation of the model is introduced, offering substantial gains in statistical power and accuracy in parameter estimates. Finally, the fourth chapter is an application of the methodology proposed in the third chapter. Specifically, the model is applied to the study of transitivity of preference in the domains of probabilistic and temporal discounting. Many instances of violations of transitivity were found at the individual level for the domain of probabilistic discounting and for the case where temporally and probabilistically discounted options were compared, with over 80% of people showing strong evidence for transitivity violations in at least one case.

Chapter 1

Are Response Time Transformations Really Beneficial?

It is common in experimental psychology to use response times as a dependent variable that indexes overall performance. Famous examples include Sternberg's demonstration of exhaustive short-term memory scanning [64] and Shepard and Metler's demonstration of mental rotation [62]. Response times are ubiquitous for assessing the effect of context on a target, and examples include classic Stroop and Simon effects [63, 66] as well as the more recent variants such as the weapons-priming tasks [16]. Overall, the analysis of response times in high-accuracy experiments has been a time-tested, popular, and fruitful approach in experimental psychology.

The question then is how to analyze such data. In this paper, we consider testing whether response time is affected by a covariate. A common example is the Stroop effect. Here, participants identify the display color of a color term (e.g., the word *RED*). The display color is either the same as the color term (e.g., the word *RED* displayed in red) or different from the color term (e.g., the word *GREEN* displayed in red). Incongruent words slow

response times, and accurately detecting this effect is a common goal.

The vast majority of researchers in these cases use familiar, well-trodden linear-model statistical tests such as t-tests, regression, and ANOVA, again with the goal of stating whether covariates affect RT or not. What we have noticed in the literature, however, is that while most researchers do so on untransformed variables, there are many examples of using transformations before analysis, most commonly either log or reciprocal transformations. Some examples of the use of a logarithmic transformation can be found in [28] and [36]; examples of reciprocal transformation can be found in [7] and [41]. Is transforming wise? Indeed, the genesis of this paper came from a friendly lunchtime disagreement among the authors about the wisdom of transformations. Over burritos, the first author expressed his discomfort about using normally distributed models for RTs, a recent approach favored by the second author [26, 27, 57, 67]. The first author recommended placing normal models on the logarithm of response time. This paper provides an assessment of whether it is wiser to perform linear-model tests on transformed or untransformed response time.

The initial recommendation to transform parameters comes from the seminal work of [13]. These authors show that with simple, univariate transforms, skewed data may be brought better in line with the normal distributional assumptions. As a result, statistical techniques that assume normally distributed residuals may be used with less worry about the effects of violating distributional assumptions. Based on this advice, some researchers recommend transformations. For example [78] makes an appeal for robust statistics including using transforms. [77] reviews the benefits of transformation, which for RT include minimizing the impact of slow-response outliers while maintaining good power, as well as a concern over the interpretation of transformed data. Skeptics of transformed variables include [51], who provides a small simulation study, and [38], who recommends generalized linear mixed models as an alternative on the basis of a case study in word naming.

In this paper, we seek to go beyond the previous assessments. Our approach relies on both an

analytic assessment of the noncentrality parameter that determines the power in statistical tests as well as a simulation check among a broad array of realistic conditions. Here, we pay particularly close attention to generate as realistic representations of RTs that we know. We opt for a shifted, one-bound diffusion model of perception [32, 49, 65] as a generative model. The parameters we use in the generative model are directly informed by trends in the empirical literature, and consequently, our results are directly applicable in real world situations.

1.1 Rationale For Transforms

To understand the dilemma facing researchers, it is helpful to review the following properties of response times. First, response times are unimodal with a skewed upper tail [39, 73]. Second, manipulations that slow the distribution tend to increase the mean and standard deviation together with relatively small effects in higher moments such as skew [58, 75]. Third, effects across conditions may be difficult to detect because there is excessive trial-by-trial variability as well as much variability across people [26]. Figure 1.1, Panel A, provides an example of these properties. Consider the two distributions that are denoted with solid lines. These two are for two experimental conditions such as *congruent* and *incongruent*. The distributions are unimodal with a large degree of skew for the upper tail. The distributions are highly similar indicating that the effect is small relative to the trial-by-trial variability. Finally. the difference between conditions is manifest in both the mean and variance.

These properties taken together indicate that a transform that disproportionately affects the long tail may be helpful in reducing skewness and stabilizing variance. Indeed, the aformentioned logarithm and reciprocal transform disproportionately reduce the long tail. Therefore, it seems reasonable that these transforms are helpful in assessing small effects. Moreover, both logarithmic and reciprocal transform have possible process interpretation.

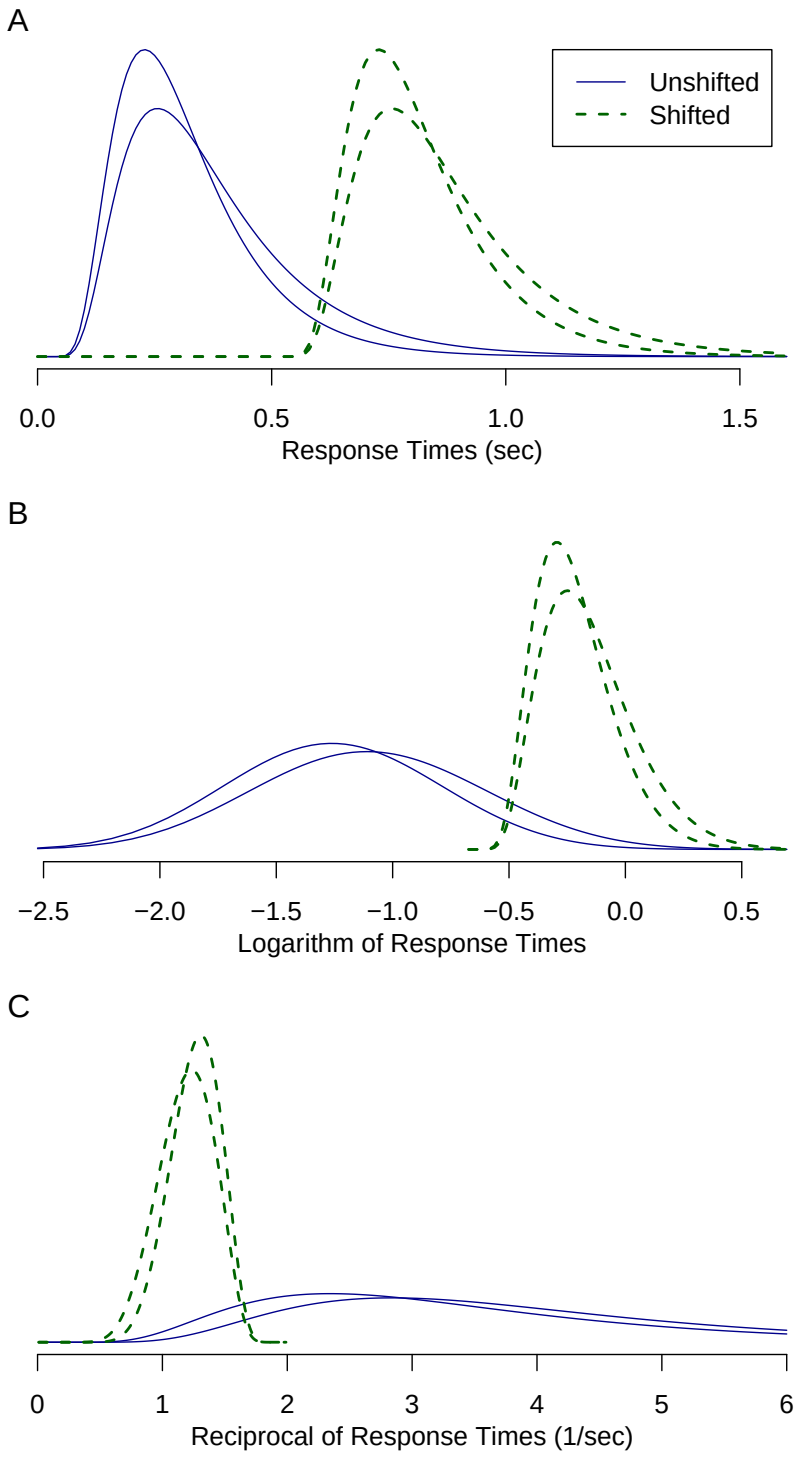

Figure 1.1: Untransformed and Transformed Response Times Distributions. There are four distributions in each panel forming two pairs of highly similar distributions. For each pair, one distribution denotes a baseline condition, the other a treatment condition. Distributions with solid lines are unshifted; those with dashed lines are highly shifted. **A**. Untransformed inverse Gaussian distributions of RT. Each pair has an effect in drift rate. **B** The distributions that result from the logarithm transforms. **C** The distributions that result from the reciprocal transform.

[71] provide four separate process scenarios that generate lognormally distributed data with most centered around multiplicative gain and attenuation from attentional filters. Likewise, [53] describe a ballistic accumulator model that naturally gives rise to response times which are the reciprocal of a truncated normal distribution. So, given this constellation of facts, it certainly seems plausible that RT transformations are beneficial in analysis.

The main goal was to see how easily we could detect an effect between two conditions. Hence, the main characteristics we assess is the performance of a conventional equal-variance t-test. For each transform, we ask which has the highest power to detect an effect with a Type I error rate maintained at the specified level. We can get a rough guideline to the first question, about power, by studying the *effect size* between two different distributions. The effect size is the difference in expected values scaled by a measure of averaged standard deviation.[1] It is a known constant for any two distributions, for example, the effect size between the solid distributions in the untransformed case is 0.34. The effect size, for fixed N directly determines the expectation of the *noncentrality parameter* of the noncentral-T, which is monotonically related to power. Hence, bigger effect sizes correspond to more power.

How does transformation affect the effect sizes between the distributions? The effect sizes for the logarithmic and reciprocal transsformed distributions are, respectively 0.33 and 0.28. Notice that these values are lower than that for the untransformed distribution. This simple effect-size analysis provides our first clue—transforming data will not increase power, and, at least for the reciprocal transform, it may even lower it. Unfortunately, while studying the effect size provides strong guidance, it is not the final answer to our question. First, it does not address the level of the test, that is, the real Type I error rate. Second, while perfectly predictive for power in the asymptotic limit, the effect-size analysis may not hold with finite samples. Nonetheless, we should be quite skeptical of any gains from transformations off the bat.

[1] Let X and Y be two distributions. The effect size is $(E(X) - E(Y))/\sqrt{.5(V(X) + V(Y))}$.

1.2 Those Pesky Shifts

There is one additional property of response times that has been overlooked by those using transforms. This property is that all response time distrubutions have a substantial shift away from zero. Shifted distributions are shown in Figure 1.1 as dashed lines, and these particular distributions have a shift value of .5 sec. Notably, there is no possibility of a response below this .5 sec value. Shifts in RT distributions are ubiquitous [39, 56]; indeed, we know of no data set with mass near zero. Figure 1.2A, from [59], provides an example. The data here come from 94 participants in lexical decision experiment from [21] where people responded in about 700ms on average. The plotted values are the shift parameter from a lognormal distributional fit as performed in [59]. The shift represents the point below which there is no mass. The thin lines are 95% credible intervals, and shifts are located far from zero for all participants.

Consideration of shifts, in our opinion, simply reflects the nature of response times themselves. After all, it is a reality of neurophysiology that the mere perception of an unanticipated stimulus is doomed to be delayed post exposure at least by the amount of time it takes for the chains of action potentials from the retina to relevant brain areas to run their course. Likewise, the same can be said for the chain of action potentials required to translate the decision into the proper button press. No wonder so many RT modeling efforts include shift parameters (e.g.,[18, 50, 74, 72]).

The presence of these shifts pose a serious challenge to the rationale for transforms. Figure 1.1, Panels B and C shows the effects of shifts. Here, the shift can dramatically change the shape and scale of the resulting transformed distribution. For example, unshifted log-transformed distributions (solid lines) are more symmetric than shifted log-transform distributions (dashed lines). Likewise, shifted reciprocal transformed distributions (dashed lines) have negative skew rather than being symmetric. Moreover, not only is the variance not

stabilized, the degree of variance is a function of the shifts, indicating that the transforms may work in some paradigms better than others.

1.3 A Simulation Approach

To understand the effects of transforming response time in realistic situations, we decided to use a simulation method. The results of a simulation are only as good as the choices made for inputs. We strove for highly realistic choices that characterized common research. We decided to use the diffusion model of perception as our base generative model. The key assumption in diffusion models is that moment-to-moment evidence accumulates gradually until there is enough total evidence to support a response. For tasks where accuracy is relatively high, say those commonly used for examining working memory, response conflict, and perceptual representations, the one-bound diffusion model is highly appropriate [58]. A visualization of the evidence-accumulation process is shown in Figure 1.2B.

The distribution of absorbtion times of the one-bound diffusion process is known as an inverse Gaussian [15] or a Wald distribution [65]. When parameterized with drift-rate (ν), bound(λ), and shift(ψ), the density may be expressed in closed-form as:

$$f(t;\psi,\nu,\lambda) = \frac{\lambda}{\sqrt{2\pi}}(t-\psi)^{-3/2}\exp\Big(-\frac{[\lambda-\nu(t-\psi)]^2}{2(t-\psi)}\Big)$$

Where $t > \psi > 0$, $\lambda > 0$, $\nu > 0$. The close form nature of this density makes it far more convenient than the two-bound diffusion model (the density is given by the convergence of an infinite sum). Moreover, sampling from the density may be performed by transforming chi-squared deviates as described in [42] and as implemented in the STATMOD package in R.

The inverse Gaussian may be parameterized in terms of shift, drift rate, and bound, as

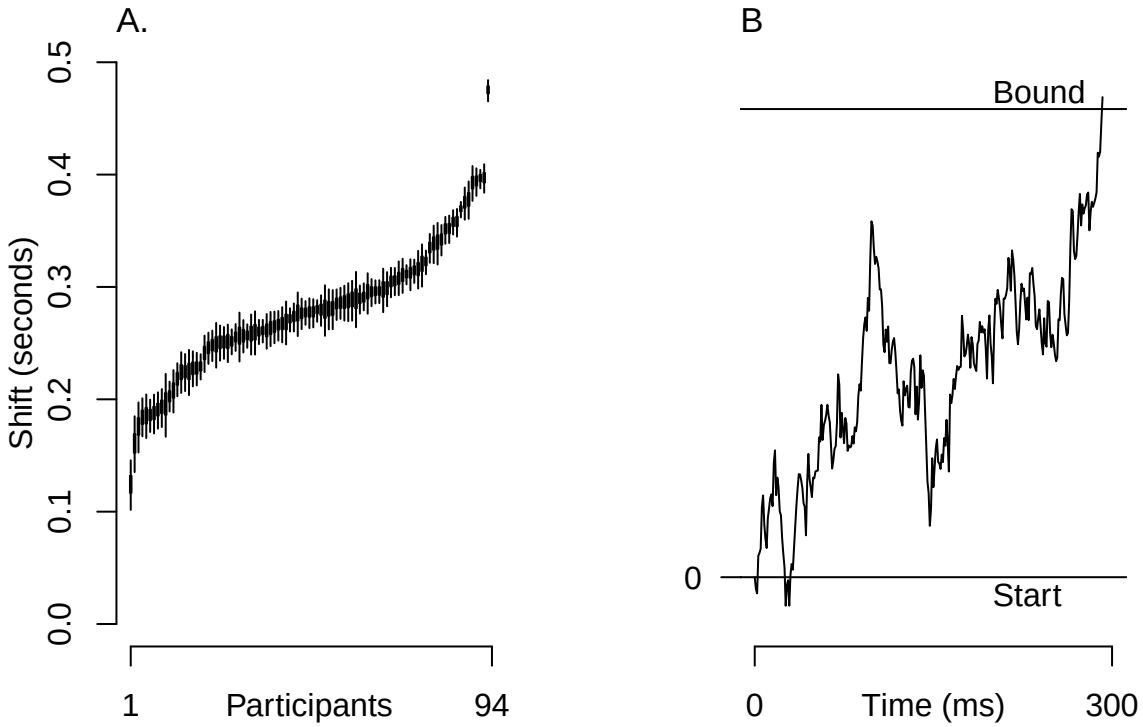

Figure 1.2: **A**. Estimated shifts from Gomez et al. (2008). **B**. An example of the accumulation in a one-bound diffusion model.

above, or, alternatively, in terms of shift, scale, and shape. For guidance about which parameterization is best for empirical phenomena, we rely on [58] who competitively tested the two different parameterzations. In one model, stimulus strength affected drift rate and bound was held constant; in the other model stimulus strength affected scale and the shape was held constant. The results were unequivocal—the constant bound model was superior. Over 90,000 observations, RT distributions showed small but detectable shape differences remarkably like those in Figure 1.1A. Given Rouder et al's demonstration we manifest the difference between our simulated conditions in drift rate.

We start with three modal response-time experiments. The first is what we term the *perceptual experiment*, and the defining characteristic is fast overall response times and small differences between conditions. We performed several variants of the fast theme, and one of these, the most realistic, serves as our *canonical version*. The canonical version of this experiment has an overall mean of 0.30 sec with a 0.02 sec difference between conditions.

Table 1.1: Parameter Values For Simulations

Name	Avg (s)	Effect (s)	Shift	Bound	Drift_1	Drift_2	Drift_0
Perception	0.30	0.02	0.15	1.20	7.50	8.57	8.04
Attention	0.65	0.05	0.35	1.20	3.69	4.36	4.03
Linguistics	1.35	0.15	0.35	1.20	1.12	1.30	1.21

Table 1.1 shows the corresponding settings of shift, bound and drift rates for power as well as a common drift rate for level. The canonical version has 60 trials per condition, and the entry is shown in the top row of Table 1.2. The remaining 8 rows show different versions made by varying the shift and the total number of trials per condition. We also simulated data from hypothetical *attention* and *linguistic* experiments. The settings for the attention experiments are designed to capture mid-level cognition experiments which take about 6/10ths of a second and have effects on the order of 50 ms. The particular settings are shown in Table 1.1, and the versions populate the middle set of rows in Table 1.2. Likewise, the linguistic set of experiments is designed to capture above-second tasks that typically have larger effects.[2] In addition to simulating data from these hypothetical experiments, we report the effect size between the distributions in each case for each transformation (Table 1.3).

1.4 Simulation Results

The results for power and level for each of the versions of each of the experiments are shown in Table 1.2. Under *Power* and *Level*, there are three columns each. The first, *RT* denotes results for untransformed distributions; the remaining two are for the logarithm and reciprocal transforms, respectively. The main results here are easy to characterize. For most cases, it matters little in power or level whether response times are transformed or

[2]The number of trials in the linguistic experiments is chosen to be greater than those in the perception and attention experiments so that the power is roughly comparable.

Table 1.2: Power from Simulation I (fixed shifts)

			Power			Level		
	N	Shift	RT	log(RT)	1/RT	RT	Log(RT)	1/RT
Perception								
Realistic	60	0.15	0.61	0.61	0.59	0.050	0.049	0.050
Realistic	20	0.15	0.24	0.24	0.24	0.050	0.049	0.050
Realistic	120	0.15	0.89	0.89	0.88	0.050	0.049	0.050
Zero	60	0.00	0.61	0.59	0.54	0.050	0.049	0.050
Zero	20	0.00	0.24	0.24	0.22	0.049	0.050	0.049
Zero	120	0.00	0.89	0.88	0.83	0.050	0.050	0.051
Large	60	0.30	0.61	0.61	0.60	0.049	0.051	0.050
Large	20	0.30	0.24	0.24	0.24	0.049	0.049	0.050
Large	120	0.30	0.89	0.89	0.89	0.050	0.050	0.049
Attention								
Realistic	60	0.35	0.50	0.50	0.48	0.049	0.049	0.051
Realistic	20	0.35	0.19	0.20	0.19	0.049	0.049	0.050
Realistic	120	0.35	0.81	0.80	0.78	0.050	0.050	0.050
Zero	60	0.00	0.50	0.48	0.39	0.051	0.050	0.050
Zero	20	0.00	0.19	0.19	0.16	0.047	0.050	0.047
Zero	120	0.00	0.81	0.77	0.66	0.050	0.048	0.051
Large	60	0.70	0.50	0.50	0.49	0.050	0.051	0.050
Large	20	0.70	0.19	0.20	0.19	0.049	0.049	0.049
Large	120	0.70	0.80	0.80	0.79	0.049	0.051	0.051
Linguistic								
Realistic	240	0.35	0.50	0.46	0.36	0.049	0.050	0.050
Realistic	80	0.35	0.20	0.18	0.15	0.049	0.050	0.049
Realistic	480	0.35	0.80	0.74	0.62	0.049	0.050	0.050
Zero	240	0.00	0.50	0.42	0.25	0.049	0.050	0.050
Zero	80	0.00	0.20	0.17	0.12	0.049	0.048	0.048
Zero	480	0.00	0.80	0.70	0.44	0.049	0.050	0.050
Large	240	0.70	0.50	0.47	0.40	0.050	0.050	0.050
Large	80	0.70	0.20	0.19	0.17	0.049	0.050	0.050
Large	480	0.70	0.80	0.76	0.68	0.050	0.051	0.049

Table 1.3: Effect Sizes

Shift	RT	log(RT)	1/RT
Perception			
0.150	0.417	0.416	0.410
0.000	0.417	0.409	0.381
0.300	0.417	0.417	0.414
Attention			
0.350	0.368	0.368	0.360
0.000	0.368	0.357	0.314
0.700	0.368	0.369	0.366
Linguistic			
0.350	0.181	0.173	0.150
0.000	0.181	0.164	0.119
0.700	0.181	0.176	0.161

not, though no transformation and log transformations seem to do marginally better than reciprocal transformations in most instances. From 1.3 we can also see this reflected in effect size. Although the t-test appears to be fairly robust in most of these cases under transformation, there is no gain that we can see in transforming the data.

To understand the full range of effects of transformation on t-test results, we plotted the receiver operating characteristic (ROC). The left column of Figure 1.3 shows the case for the canonical versions of each modal experiment. The x-axis of each ROC is the nominal level for a test, or, more conventionally, the α setting. The y-axis is the rejection rate, which describes both the power of a test (when there are true differences between conditions) and the level of a test (when there are no differences between conditions). These ROCs show that in all cases, the level is fairly close to the nominal value, which is heartening. Moreover, the overlap of all curves in the perception and attention experiments show how robust the t-test is to the transformations for these ranges of observations. The outlier is the linguistics experiments, but here, contrary to conventional wisdom, the highest power is attained with the untransformed distributions. Therefore, we conclude that there is nothing to gain by

transforming the data.

1.5 Variation in Shift

One feature of the above simulation is that the parameter values were held constant across all replications of a simulation. We therefore call these the fixed parameter simulations. It is common, however, to add variability to the parameters of the diffusion model in application. For example, [52] added variablity to the drift rate, bound, and shift to more faithfully model their data. [61] added exponential variability in the shift, but we believe that adding a small-degree of normal variation is more realistic. As shown in 1.4, this addition has the effect of making the decrease of the left tail more gentle.

We ran the simulations with a small degree of variability in the shift parameter. The shift used to generate each simulated reaction time was drawn from a normal, centered on the shift values indicated in Table 1.5, and with a standard deviation of 15% of the mean shift value. Power and level results are shown in Table 1.5, effect sizes in 1.4 and ROCs are shown in the right column of Figure 1.3. In the tables, there are no entries for zero shift in this case because we did not wish to entertain negative shift values.

In each of the cases effect sizes appear marginally better untransformed, closely followed by the log transform, and then by the reciprocal transform. This pattern is also for the most part reflected in our power simulations. Again, the results of the level simulations seem to be more or less as they should be in all cases. The ROC curves demonstrate more of the same, though with a random shift they seem to favor untransformed data ever so slightly more.

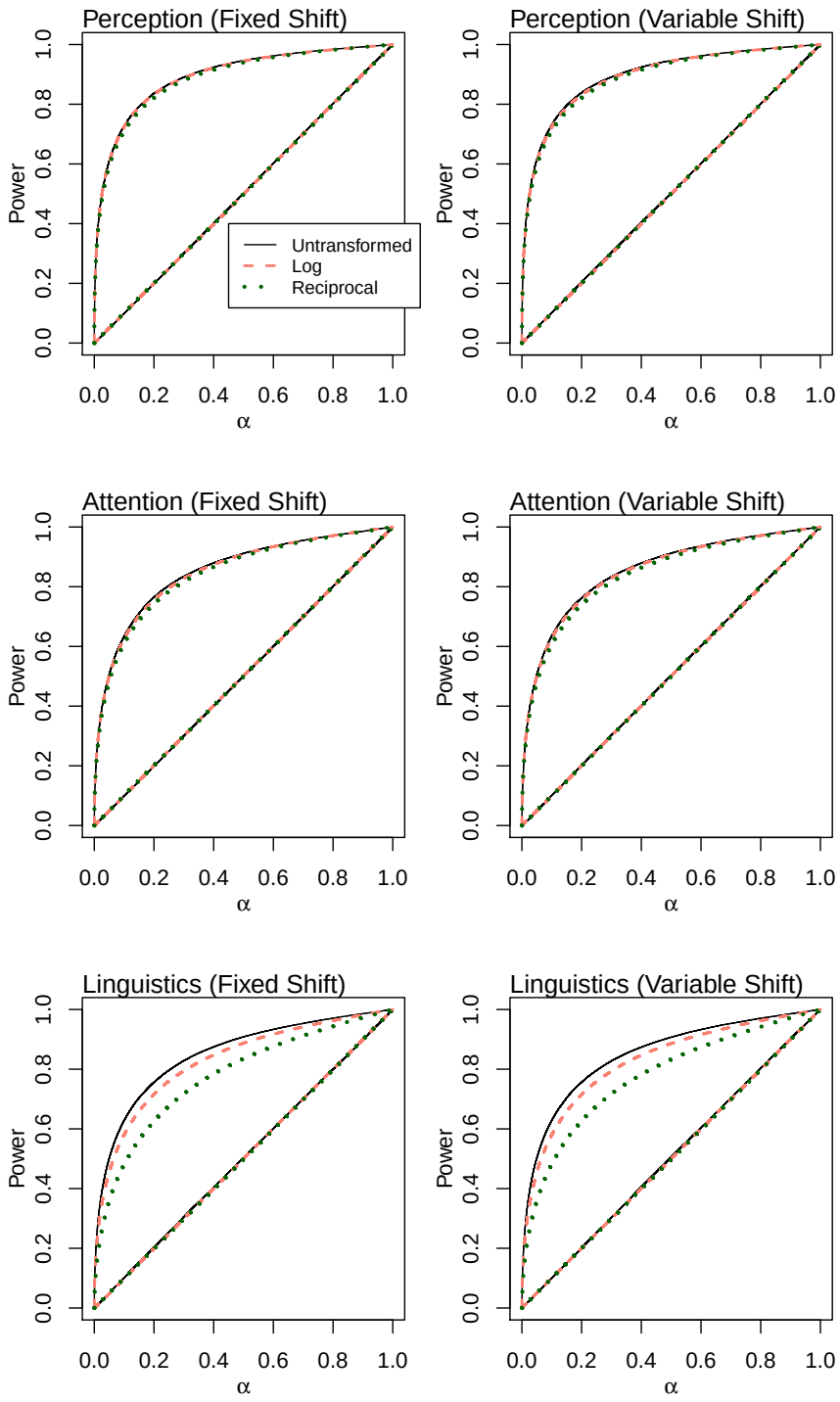

Figure 1.3: ROC plots for each of the canonical cases. Left column: fixed shift, right column: normally distributed shift.

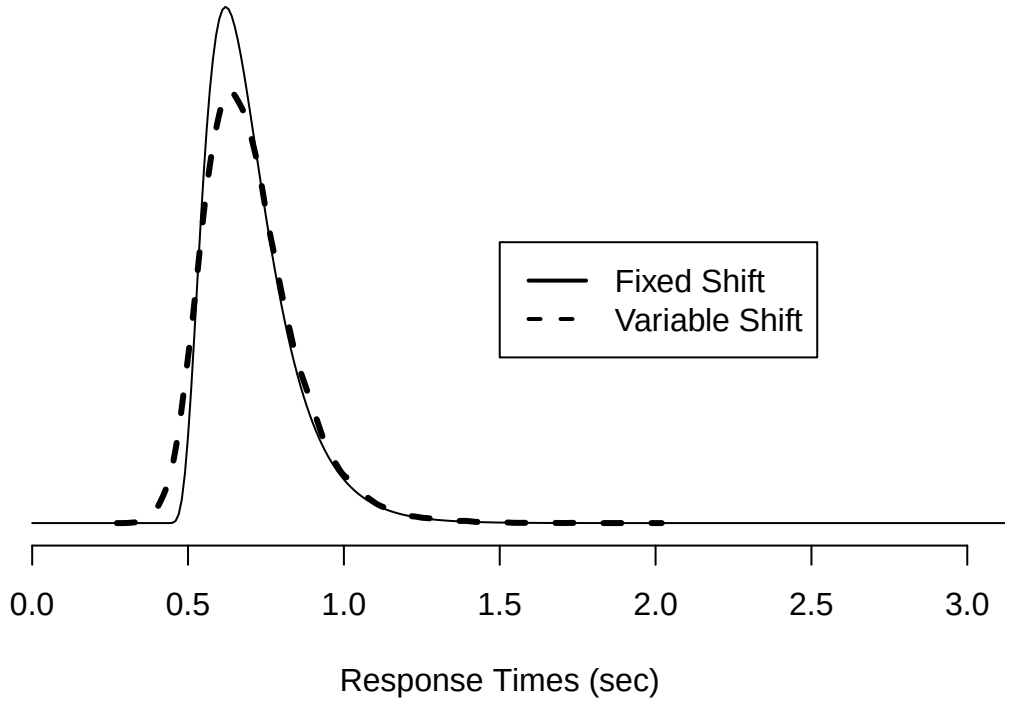

Figure 1.4: The effect of adding a small degree of variability to the shift parameter is to make the left tail more gentle.

Table 1.4: Effect Sizes, Variable Shift

Shift	RT	log(RT)	1/RT
Perception			
0.150	0.380	0.374	0.360
0.300	0.311	0.303	0.291
Attention			
0.350	0.330	0.322	0.306
0.700	0.279	0.269	0.253
Linguistic			
0.350	0.177	0.169	0.147
0.700	0.176	0.170	0.154

Table 1.5: Power from Simulation II (random shifts)

			Power			Level		
	N	Shift	RT	log(RT)	1/RT	RT	log(RT)	1/RT
Perception								
Realistic	60	0.15	0.53	0.51	0.49	0.049	0.049	0.049
Realistic	20	0.15	0.21	0.20	0.19	0.049	0.051	0.050
Realistic	120	0.15	0.82	0.81	0.78	0.049	0.050	0.050
Large	60	0.30	0.37	0.36	0.34	0.051	0.049	0.049
Large	20	0.30	0.15	0.15	0.14	0.049	0.051	0.050
Large	120	0.30	0.64	0.62	0.59	0.049	0.050	0.049
Attention								
Realistic	60	0.35	0.45	0.43	0.40	0.050	0.050	0.050
Realistic	20	0.35	0.17	0.17	0.16	0.049	0.052	0.049
Realistic	120	0.35	0.75	0.72	0.68	0.050	0.050	0.049
Large	60	0.70	0.34	0.33	0.30	0.050	0.049	0.050
Large	20	0.70	0.14	0.13	0.13	0.049	0.050	0.050
Large	120	0.70	0.60	0.57	0.53	0.049	0.049	0.051
Linguistic								
Realistic	240	0.35	0.50	0.45	0.35	0.049	0.050	0.050
Realistic	80	0.35	0.20	0.18	0.15	0.050	0.050	0.050
Realistic	480	0.35	0.79	0.74	0.61	0.050	0.050	0.049
Large	240	0.70	0.50	0.45	0.37	0.050	0.050	0.050
Large	80	0.70	0.19	0.18	0.16	0.049	0.051	0.050
Large	480	0.70	0.79	0.74	0.64	0.049	0.049	0.051

1.6 Discussion

In this paper, we consider the effects of transforming response times in realistic cases through analysis of the effect sizes of distributions and through simulation. Our results are clear. Transforming variables offers no more power nor any better level control than not transforming. In specific cases, certain transforms, notably the reciprocal transform, provided for lower power. Thus, we see no reason to transform response times in establishing effects.

The remaining question then is about the generality of the results. We think they hold fairly broadly when the goal of the researcher is to establish effects through linear model analysis. For example, we would expect no increased power from transforming if our goal was to detect whether RT covaried with variables such as word frequency. Likewise, we see no reason to transform variables in ANOVA analyses where the goal is to test main effects and interactions. Simply put, if the goal is to establish the presence or absence of effects, then transformation holds no advantages that we can identify.

A corollary to this result is that normal models may be used for response time when the goal is to establish nominal or ordinal relationships. Consider the goals of [26] who asked what they call the "does everybody" question. We know that *on average* people respond more quickly to congruent items than incongruent ones in the Stroop task. Does this ordering, congruent-faster-than-incongruent, hold for all people, or, alternatively, are their people who truly have no Stroop effect or have a pathology where they truly respond more quickly to incongruent items? Given that their questions are about detecting the presence and direction of effects, the results here indicate that the normal is not problematic. Indeed, [67] use simulations to show normal models of response time distributions are perfectly reasonable for such ordinal questions.

Researchers interested in modeling the functional relationship between RT and covariates—say whether RT goes as a power function or an exponential function of a covariate—do

need more accurate models of RTs. [60] showed that a main concern is modeling the shifts. A failure to do when shifts are present leads to systematic distortions of the functional relationship between RT and covariates. The presence of shifts that vary across people is a strong challenge because models in the linear mixed model and generalized linear mixed model family are unable to capture individual-specific shifts. Fortunately, researchers do have an option. Bayesian hierarchical models are perfectly suited for the challenge, and recent developments in general purpose software such as stan [14] and JAGS [47] make analysis with these models convenient for nonexpert researchers. There are now several excellent sources for guidance on hierarchical models in cognition including [31] and [35].

Chapter 2

Hierarchical Paired Comparison Modeling, A Cultural Consensus Theory Approach

2.1 Introduction

Paired comparisons are ubiquitous within Psychological science as a means to gain information about people's relative judgments regarding a set of stimuli. Because of the wide usage of paired comparison studies, a principled analysis of such data in a manner that incorporates information gained about differences between both people's response patterns and individual item effects can be useful. Developing an appropriate model to analyze such data is the focus of this paper.

In other item response domains, Cultural Consensus Theory (CCT) has proved useful. CCT is a set of models developed by [55] that analyze response data by means of assuming one or more sets of latent ground truths common among a group of people. In CCT, what has

traditionally been referred to as "ground truth" need not be a reflection of reality, rather it is a reflection of group opinion, although CCT has proven useful also for wisdom of the crowds purposes in some instances as well (e.g. [1]). CCT accounts for individual differences in propensity toward answering according to the ground truth, and also accounts for differences in difficulty of evaluating certain items according to the consensus ground truth. CCT uses this information to assess a consensus ground truth while simultaneously providing measures for item difficulty and subject agreement. CCT item response models have been developed for binary response data [4], multiple choice [55], continuous closed interval responses [6], and ordinal data [2]. For an overview of CCT, see [5].

Thus far, there has not been a CCT model designed for paired comparisons. This paper introduces such a model. Given a complete set of paired comparisons between a group of items (e.g. comparing preference), the goal is to measure consensus attitudes from people's choices. In the spirit of CCT, our model also identifies which people's responses tend to be the closest to consensus, and which items are particularly difficult to assess, and take these things into account in the measurement.

The backbone cognitive model behind our Consensus Paired-Comparison Model (CPCM) is the Thurstonian model [68]. The assumptions of the Thurstonian model are similar in nature to signal detection theory in that the subjective experience of stimuli is assumed to follow a Gaussian distribution on a latent scale. In application to subjective value judgments, for example, a stimuli could be thought to have its appraised latent value judgments mean-centered at a particular value, and distributed normally. For paired comparisons, the assumption is that each option is evaluated according to its own Gaussian distribution on each trial and the higher-valued option is the one that is selected. Under the assumption of equal-variance Gaussian distributions (referred to as the case V model), the Thurstonian model reduces according to an ordinary probit model. Allowing for separate variances from independent Gaussians gives the Case III model, which is what is used in CPCM, where the

probability of choosing option A over option B is

$$P_{A \succ B} = \Phi \left(\frac{V_A - V_B}{\sqrt{\sigma_A^2 + \sigma_B^2}} \right) \tag{2.1}$$

Here Φ denotes the cumulative distribution function of the standard normal distribution, V_A and V_B denote the mean of the Gaussian distributions of value assessment, whereas σ_A^2 and σ_B^2 denote their respective variances.

Researchers who utilize the models that we introduce will likely be most interested in finding consensus values for the Vs in equation 1 for each of the options in the set of items. Since we utilize a Bayesian approach, we end up with posterior distributions for these values. Below is a set of inferred posterior distributions of these values measured using the WCPCM (described later) from a data-set in which participants were asked to compare different professions according to which mean salary they believe to be higher. Note that the scale used is not in monetary amounts, rather it is according to the Thurstonian scale which should have a monotonic relationship with salary estimates.

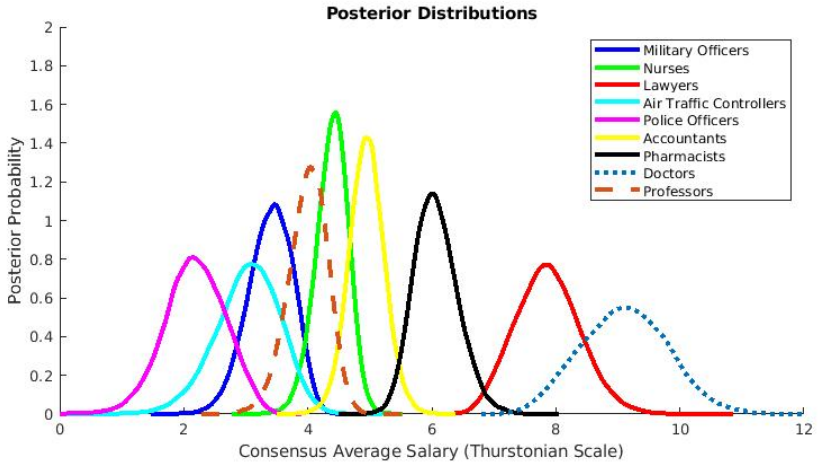

Figure 2.1: Posterior Distributions from the WCPCM for the careers dataset

In this paper we discuss two separate models. The first model, referred to as the Strong Consensus Paired Comparison Model (SCPCM), reflects the essence of previous CCT models in that it assumes individuals always answer items from the same consensus "answer key", with each response carrying the possibility of independent error. We overview the model, strengths and shortcomings of this approach in the following sections. Later we will discuss an expansion of this model that relaxes these assumptions which we will refer to as the Weak Consensus model.

2.2 Model 1: Strong Consensus

The Strong CPCM (SCPCM) model is described according to the following axioms:

Axiom 1: (Common Truth). For group c of the C cultural groups, there is a row vector T_c of latent values on the real line for each of the M items. T is a $C \times M$ matrix.

Axiom 2: (Thurstonian Paired Comparisons) The probability for subject i picking any item A over any item B is given by $P_{i,A \succ B} = \Phi\left(\frac{T_{g_i A} - T_{g_i B}}{\sqrt{\kappa_i (s_{g_i A} + s_{g_i B})}}\right)$. Where Φ is the inverse probit function, and g_i indicates which cultural group c subject i is in. Note that this is equivalent to the Thurstonian case III model(Thurstone, 1927), where each Gaussian for item k is centered at $T_{g_i k}$ and has variance equal to $\kappa_i s_{g_i k}$.

Here, κ_i can be seen as subject i's tendency to diverge from the group consensus, with larger values indicating more divergence whereas s_{ck} can be seen as a measure of the level of inconsistency of people in group c's evaluation of item k, with larger values indicating less consistency.

In this strong consensus version of the model, it is assumed that everyone in the same group has the same latent value for each item, but individual subjects have different variances in their assessments of an items value.

Built into the assumptions of the case III Thurstonian model is weak stochastic transitivity on the individual level, and in the case of the SCPCM on the group level too, since all in group c share the same latent values T_c. What this means is that if $P_{A\succ B} \geq \frac{1}{2}$ and $P_{B\succ C} \geq \frac{1}{2}$, then $P_{A\succ C} \geq \frac{1}{2}$ for any of these probabilities from any individual in the same group c.

2.2.1 Hierarchical Bayesian Parameter Estimation

We use a hierarchical Bayesian approach to estimate the parameters in the model. Hierarchical Bayesian modeling involves specifying stronger distributional assumptions regarding not only the data itself, also the parameters involved in generating the data, yielding provably better estimates [19]. In many cases one might opt to assume one single cultural group ($C = 1$), as was done in the analyses in this paper. For other applications when it is suspected that there is more than one cultural group ($C > 1$), the following specification could be utilized.

$$g_i \sim \text{Categorical}(\pi) \tag{2.2}$$

$$\pi \sim \text{Dirichlet}(L) \tag{2.3}$$

Here g_i can take any integer from 1 to C. π in this case is a C dimensional vector of probabilities corresponding to the probability of a person being in each group, and is specified to come from a Dirichlet distribution parameterized by L. L can be set to equal a C dimensional vector of ones, which is essentially a C dimensional extension of the Uniform distribution. Alternatively, one can set L such that it gives the most weight to the first cultural group and assigns descending weight to the succeeding groups.

The consensus value T for item k and cultural group c is given the following prior distribution:

$$T_{ck} \sim \mathrm{N}(5, .25)^1 \tag{2.4}$$

In this case, centering at 5 is completely arbitrary, chosen entirely for the sake of yielding values usually between 0 and 10. Setting the precision (inverse variance) at .25 however is chosen so that the model can be properly identified.

Now both subject and item specific components of the Thurstonian variance are given lognormal priors, centered at 0 with uninformative gamma distributed hyperpriors for the precision:

$$log(\kappa_i) \sim N(0, \tau_\kappa) \tag{2.5}$$

$$log(s_{ck}) \sim N(0, \tau_s) \tag{2.6}$$

$$\tau_\kappa \sim gamma(.01, .01) \tag{2.7}$$

$$\tau_s \sim gamma(.01, .01) \tag{2.8}$$

Recovery Analysis for SCPCM

Data was simulated from SCPCM for different numbers of subjects and items for a recovery analysis. Parameters used were drawn in accordance with the hierarchical specification, with both τ_κ and τ_s set to 1. From these results we see that the model is successful in recovery, with better results when there are more subjects and items. Under fewer subjects and items, results are qualitatively accurate but are prone to some shrinkage from the hierarchical structure.

[1]This is a mean-precision parameterization of the normal distribution

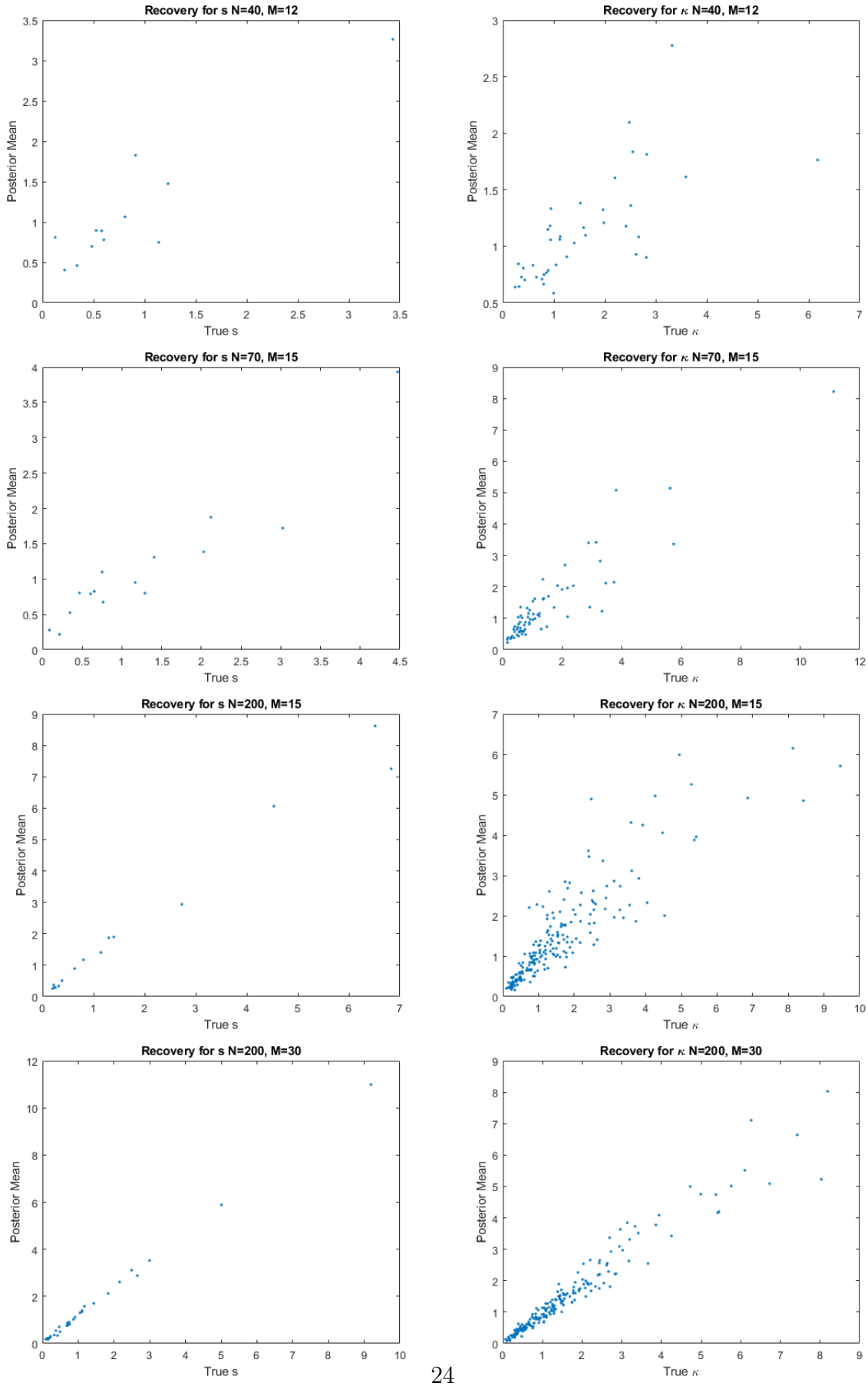

Figure 2.2: Recovery analysis for SCPCM

2.3 Data Sets

We applied our models to three different datasets, each with 9-15 different items. In every case, all participants answered which item they choose between every possible pair of the set of items. For each data set, an attempt was made to avoid repeating the same items too close to one another to avoid memory effects.

2.3.1 Occupation Salaries

40 subjects were asked to compare the salaries of 9 different occupations, such as pharmacist and police officer. Data was collected after a test in an undergraduate Psychology course, giving students the opportunity to fill out the survey in exchange for bonus points.

2.3.2 Car Prices

66 participants were shown pairs of pictures of 10 cars, with their make/model. Participants were asked which car they thought was more expensive of each pair. The study was conducted on computers, and participants were recruited through the schools online subject pool. Participants received credits through the subject pool that can usually be used for extra credit in some undergraduate Psychology courses.

2.3.3 Cheerfulness of Paintings

The same 66 participants that responded to Car Prices were also asked to compare which painting seemed more cheerful for each pair out of 15 different paintings.

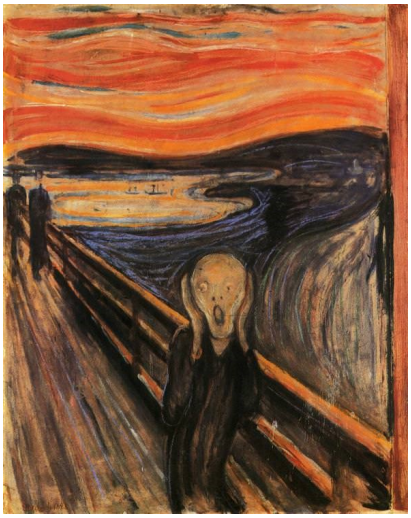

Figure 2.3: Two of the paintings used. Left: Bedtime Aviation by Rob Gonsalves, Right: The Scream by Edvard Munch

2.4 Posterior Predictive Tests

We utilized posterior predictive tests to check whether our models were effective at capturing important nuances in the data, using the same approach as in [20].

The idea behind posterior predictive tests is as follows. Since MCMC sampling samples from the joint posterior distribution of all parameters, we effectively get a set of possible parameter values for each iteration. We used each sample to construct a new simulated paired comparison data set. The simulated data sets allow for an extraction of a distribution for certain statistics of interest that we can compare with the observed statistics to check whether the model successfully captures the necessary information. Although these tests are traditionally referred to as posterior predictive tests, it should be noted that these distributions aren't actually predictions of new, untrained data, rather they represent the data you would expect to see if the model specification reflected the true generative stochastic process. Thus, it can be used as a tool to diagnose whether the model assumptions are

inconsistent with observations, but not as tool to protect against overfitting.

The following three subsections describe the different posterior predictive tests we employed to evaluate whether our models were capturing what we wanted.

2.4.1 Scree Smears

Scree plots are constructed by doing a eigenvalue decomposition on the correlation matrix of the subject by item response matrix and plotting the values of the highest eigenvalues in a descending fashion. Previous CCT models have utilized Scree Plots primarily as a tool to detect the number of consensus answer keys [3], in this case corresponding to the value of C under axiom 1. When C is one, one can expect a very high first eigenvalue followed by a sharp drop and quick plateau, whereas when C is two or more the sharp drop generally comes later. Besides the number of cultural groups, the general degree of consensus and similarity of people's responses can also affect the scree plot, with less agreement leading to scree plots that are slower to decline.

To test whether the model used is plausible, we compare the scree plots of the real data set with those of the simulated sets. We do this with a scree plot of the raw data, as well as using spearman rank-order correlation from the implied ranking through summation of all the times an item was preferred by each individual according to their responses.

2.4.2 Violations of Transitivity

If during the experiment a subject picked item A over B, B over C, and C over A, that subject has violated transitivity. From the simulated data sets, we can sum the total number of intransitivities over every triplet of items for every subject, and compare to the actual

observed number. This measure can help inform us of whether the model reflects behavior on an individual level well.

2.4.3 Between-Subject Transitivity Violations

The previous check was effective in assessing the model's accuracy accounting for individual behaviors. But what about between people? This time, a violation of transitivity was counted if Subject A picked item i over j, B picked item j over k, and C picked item k over i. All possible combinations and permutations of 3 people and 3 items were examined and summed over to calculate the total between person violations of transitivity. This is an interesting statistic to look at because if there is a high level of consensus on a set of items one would expect a low amount of between-subject transitivity violations, and vice versa.

2.4.4 Posterior Predictive Results for the Strong Model

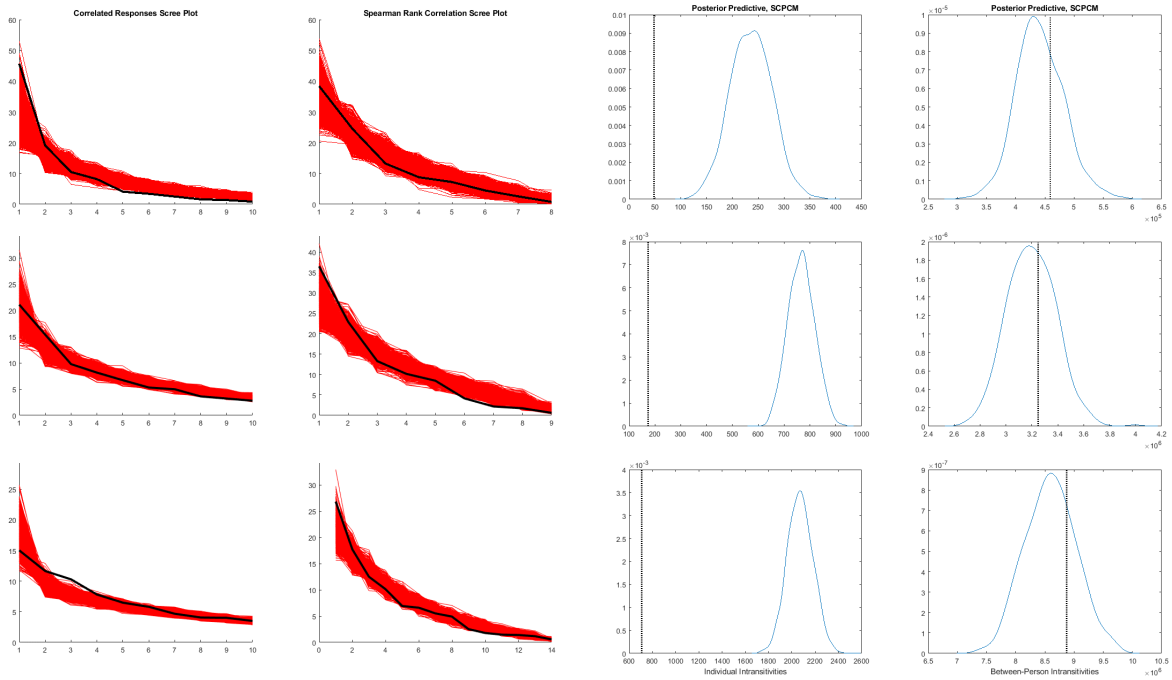

Figure 2.4: Posterior Predictive plots for the SCPCM fit to the occupations (first row), cars (second row), and paintings (third row) datasets.

In these simulations we see that the model does a decent job with most of the posterior predictive statistics, but it consistently drastically overestimates the number of individual transitivity violations that should occur. Apart from that, only the Pearson scree plot for the painting cheerfulness task seems a bit off. While this model seems to be mostly consistent with response phenomena at a larger group level, it could use some improvement for accounting for individual response behavior.

The reasoning behind this discrepancy was considered. The Strong Consensus Model assumes that everybody has the same underlying ground truth, and each time people evaluate the value of an item they are drawing from a distribution centered in the same place as it is for everyone else. This assumption implies that every time someone makes a decision that greatly diverges from consensus, they would have been just as likely to diverge from consensus in the opposite direction. In other words, the model treats divergence from consensus to

be the same as a lack of self-consistency, this seems like the likely source of the model's inconsistency with observations. To account for this inconsistency we developed the Weak Consensus Paired Comparison Model (WCPCM).

2.5 Axioms of the Weak Consensus Model

The weak consensus model differs from the strong consensus model in that it assumes that people do not necessarily share precisely the same latent opinion on the values of the items, but people in the same group will have latent opinions on the values of the items that are similar to one another. In essence, one can think of the WCPCM as the SCPCM with an extra layer between group consensus and deliberation, accounting for underlying disagreement.

Axiom 1: (Common Truth). For group c of the C cultural groups, there is a row vector T_c of latent values on the real line for each of the M items. T is a $C \times M$ matrix.

Axiom 2: (Individual Latent Item Values). Subject i has a fixed latent item value for item k given by $Y_{ik} = T_{g_i k} + \epsilon_{ik}$, where $\epsilon_{ik} \sim N(0, \frac{1}{E_i \lambda_{g_i k}})$ (precision notation) and g_i indicates which cultural group c subject i is in.

Axiom 3: (Thurstonian Paired Comparisons) The probability for subject i picking any item A over any item B is given by $P_{A \succ B} = \Phi\left(\frac{Y_{iA} - Y_{iB}}{\sqrt{\kappa_i (s_{g_i A} + s_{g_i B})}}\right)$. Where Φ is the inverse probit function . Note that this is equivalent to the Thurstonian case III model(Thurstone, 1927), where each Gaussian for item k is centered at Y_{ik} and has variance equal to $\kappa_i s_{g_i k}$.

It should be noted that E_i can be seen as a measure of subject i's tendency to view items as having a value close to his/her cultural group g_i, with small E_is denoting a tendency of being closer to their cultural group. Likewise, λ_{ck} can be seen as the variability in people's latent item values across different people in cultural group c for item k.

Similarly, κ_i can be viewed as a measure of individual i's consistency in their assessment of item value for the same item, with lower values indicating more consistency. s_{g_ik} Can be viewed as the tendency of item k in being evaluated consistently for cultural group g_i, with lower values indicating more consistency in evaluation for that item within an individual's evaluations.

This time, weak stochastic transitivity is again followed as a consequence of the case III Thurstonian model, but only on the individual level, since individuals do have different latent values for the items.

2.5.1 Hierarchical Bayesian Parameter Estimation

The Hierarchical specification of the WCPCM is identical to the SCPCM, only with the addition of the λ and E terms which are given the same sort of hyperprior as the κ and s terms:

$$log(\lambda_{ck}) \sim N(0, \tau_\lambda) \tag{2.9}$$

$$log(E_i) \sim N(0, \tau_E) \tag{2.10}$$

$$\tau_\lambda \sim gamma(.01, .01) \tag{2.11}$$

$$\tau_E \sim gamma(.01, .01) \tag{2.12}$$

Recovery Analysis for Weak Model

Similar to before, data was simulated according to the model with all the various τs set to 1. We can see here that despite not being as accurate in terms of estimating the exact raw parameters (being a more complex model), the qualitative results are approximately right and get better with more participants and more items.

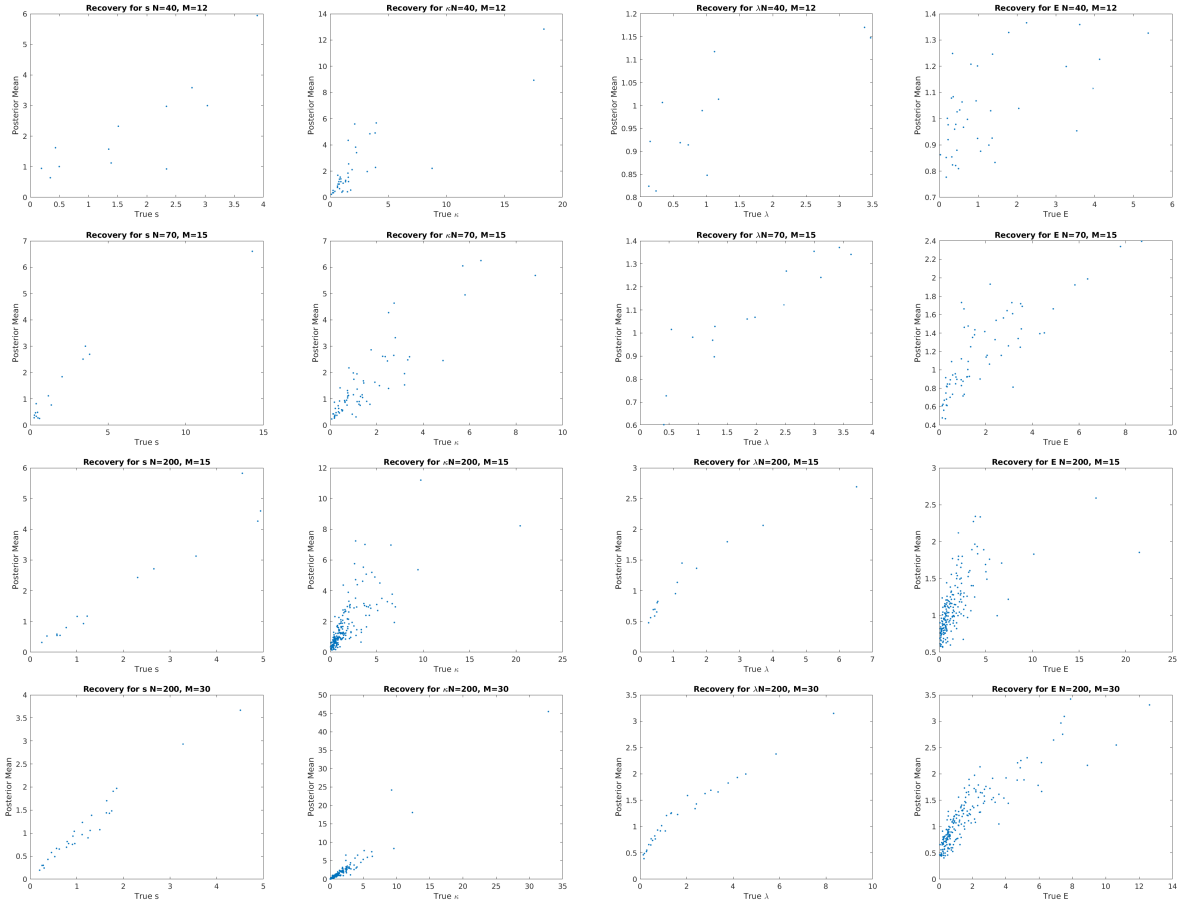

Figure 2.5: Recovery analysis for WCPCM

Do Items Really Vary in Individual Consistency?

It's not immediately apparent that $s_{g_i k}$ is really a necessary parameter. One could set them all equal to 1 and then that would only leave a measure for individual self-consistency and it might work just as well. A Bayes Factor was used to test this for the three data sets, giving

50 percent prior probability to all of the s_{g_ik} s equaling 1, and 50 percent prior probability for the full weak consensus model. The Bayes Factors came out just barely in favor of the full weak consensus model for the paintings and occupations, which came out to be 1.01 and 1.56 respectively. For the car prices set however, evidence for the full weak consensus model was very strong, with a Bayes Factor of 93.3. You can see the results summarized in the table below.

Model	Price	WC Mean Value	SC Mean Value	WC Disagreement	WC Inconsistency	SC "Difficulty"
Toyota Camry LE	22520	3.0217	4.1075	0.7482	0.284	0.5485
Porsche 718 Boxter	58450	9.2753	6.767	0.9567	1.8489	0.6046
Acura ILX	29065	3.8795	4.5772	1.0312	0.143	0.5085
Subaru Outback	25053	3.7294	4.5229	1.3327	0.4239	0.4962
Jeep Wrangler	25090	5.2976	5.1906	1.3729	0.4258	0.5003
Audi R8	166150	8.7639	6.435	1.1779	2.4979	0.5768
Ford F150 Regular Cab 2WD 6.5’	26977	3.8854	4.5368	1.5147	0.4178	0.5339
Mazda MAZDA3	17862	3.7915	4.5675	1.072	0.2136	0.4992
Honda Civic	19267	3.0383	4.1539	0.7376	0.6615	0.5645
Land Rover Discovery	53085	6.1368	5.5125	1.2351	0.5491	0.5047

Figure 2.6: Results for the car dataset

As we can see there were certain cars that people were substantially more inconsistent with than others. Interestingly this was strongest for the two most expensive cars, both of which happened to be sports cars. What’s more, we see with the Strong consensus model almost no difference in difficulty in assessment, which might be interpreted to indicate that there is about the same level of agreement about each car, but the results from the Weak Consensus model tell a different story about the varying levels of agreement. From the Weak consensus model we can see that there was more disagreement surrounding the prices of the pickups and SUVs, while the more generic cars like the Toyota Camry and Honda Civic had more

agreement surrounding their prices. This would go unnoticed if we stuck to the strong consensus model.

2.5.2 Posterior Predictive Tests for the Weak Consensus Model

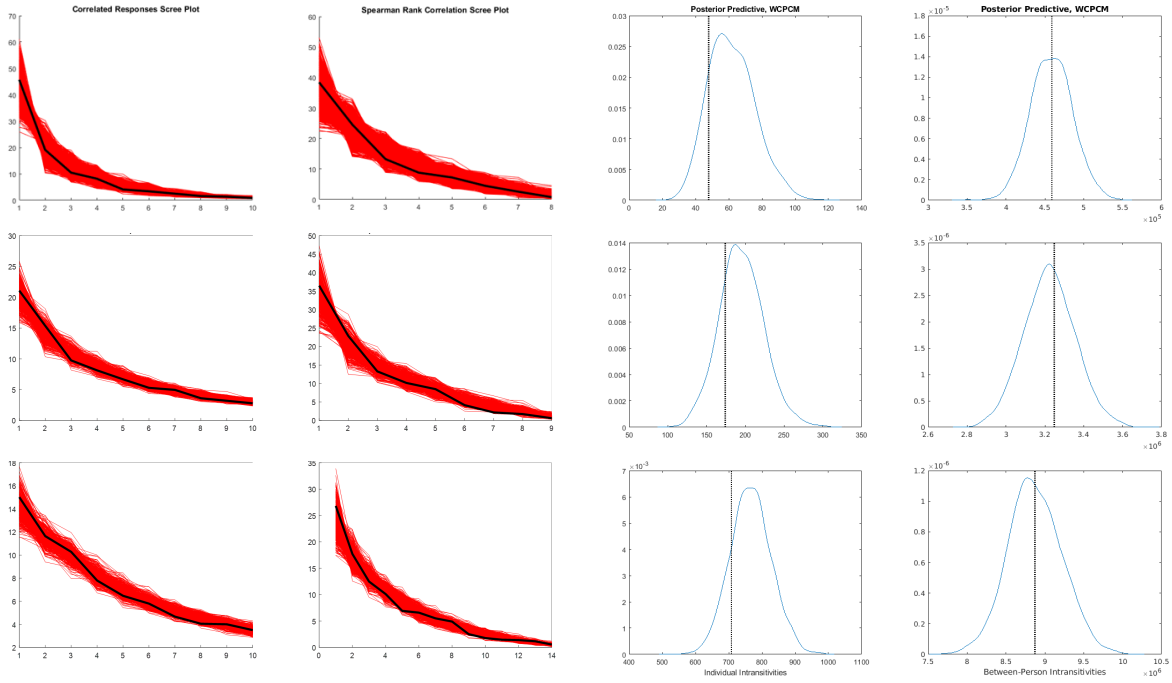

Figure 2.7: Posterior Predictive plots for the WCPCM fit to the occupations (first row), cars (second row), and paintings (third row) datasets.

It can be seen that every issue with the posterior predictive checks that was present in the strong consensus model is not present in the weak consensus model. We see true statistics represented in reasonable locations in the posterior predictive distributions.

2.6 Discussion

The SCPCM and WCPCM presented in this paper are able to serve as effective tools in the analysis of paired comparison data, quantifying the consensus values of a set of items

on a Thurstonian scale. While the SCPCM makes stronger assumptions that might be less realistic (as illustrated in posterior predictive checks), its simplicity and better identifiability still make it worthy of use for some applications. The WCPCM's relaxation of assumptions and inclusion of additional measures can make it a more attractive choice when there is a sufficient number of items and participants. Indeed, as we saw that in the car prices dataset, the WCPCM picked up on properties regarding levels of agreement on prices between different kinds of automobiles that were non-emergent from the SCPCM analysis, along with tendencies regarding consistency in item analysis. By treating people's variability in assessment in their own choices as separate from their tendency to respond in accordance with group consensus, we are able to uncover the full story more.

A limitation of these models is that they fail to account for any potential pairwise context effects since they utilize the case III Thurstonian model. This is a potential avenue for an extension of the currently proposed approach. For example, one might include additive terms to either the T or the Y terms that are only added in the presence of specific alternatives to test for context effects under this framework. The power and robustness of such approaches to testing context effects under this framework are worthy of exploration. It may also be feasible to utilize the Case I Thurstonian model for these purposes, which models latent appraisals from paired comparisons as coming from multivariate normal distributions as opposed to two independent gaussians.

As with other CCT models, it would be worthwhile to dedicate more formal exploration into their potential for wisdom of the crowds applications. The SCPCM as introduced here shares many similarities with [34], which was designed for these purposes. While [34] used rank-order data and the SCPCM and WCPCM use paired-comparison data, it might be interesting to compare these models in their usefulness for wisdom of the crowds applications. Preliminary work into using a CCT approach for wisdom of the crowds applications such as in [1] has shown promising results, thus it is worthy of further exploration.

Chapter 3

The Individual True and Error Model: Getting the Most out of Limited Data

3.1 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. 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. [70], [29], [54]). [8] 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 look only at marginal choice probabilities, that is the probabilities of responses to individual binary choices, rather than examining co-occurrence of choices. As pointed out by [9], these assumptions can lead to wrong conclusions about people's true underlying sets of binary preferences. If we take for example transitivity of preference, 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 [54]. Conversely, it is also theoretically possible for people to be fully intransitive at any point in time but if they were to reverse their preference ordering throughout the experiment the marginal choice probabilities can still give the appearance of adherence to stochastic transitivity (e.g. if 80% of the time $a \succ b, b \succ c$, and $c \succ a$ and 20% of the time $b \succ a, c \succ b$, and $a \succ c$).

The concern with independence has motivated the development of true-and-error (TE) models which do not presume full response independence. TE models originally evolved out of the approach in [37]. 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 .5 for each binary choice [10]. 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 (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, or individual data, where each participant completes multiple blocks.

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 require multiple sessions each lasting an hour or more to be necessary to achieve the statistical power necessary to reject a set of constraints. Thus, it is of particular interest to researchers utilizing 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 advocated for in [10] and [12] 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 [33], the focus of this paper is on patterns of preferences between 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 true-and-error models.

3.2 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 analyzes. When utilizing 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.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{3.1}$$

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. 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:

$$P(T_{i,m} = j) = p_{i,j} \tag{3.2}$$

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:

$$P(f_{i,m}|T_{i,m}) = \prod_{k=1}^{3} I(A_{f_{i,m},k} = A_{T_{i,m},k})[1 - e_{i,k}] + [1 - I(A_{f_{i,m},k} = A_{T_{i,m},k})]e_{i,k} \tag{3.3}$$

and

$$P(g_{i,m}|T_{i,m}) = \prod_{k=1}^{3} I(A_{g_{i,m},k} = A_{T_{i,m},k})[1 - e_{i,k}] + [1 - I(A_{g_{i,m},k} = A_{T_{i,m},k})]e_{i,k} \tag{3.4}$$

Where I denotes an indicator taking values of 0 if the statement inside is incorrect, and 1 if correct.

It should be noted that the 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:

$$P(f_{i,m}, g_{i,m}) = \sum_{j=1}^{8} p_{i,j} P(f_{i,m}|T_{i,m} = j) P(g_{i,m}|T_{i,m} = j) \tag{3.5}$$

3.3 Shortcomings and Improvement to the Present Frequentist Approach

For the sake of fitting and testing TE models, [10] 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 [46] is advised 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 was utilized in [12].

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 here 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 is 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, with a true type I error rate far lower than the nominal α level, and p-values that 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 suppose to feature a Chi-Square null distribution with degrees of freedom equal to the difference in number of outcome probabilities fixed, while in the case of the TE model we are fixing model parameters that have some effect on potentially all outcome probabilities. Thus, it turns out that this reduction in degrees of freedom is costing us greatly all while not fulfilling its 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 [45]. The famous Neyman Pearson Lemma introduced in [45] proves that the Likelihood Ratio Test is the single most powerful test. While the distribution of the test statistic (2 times the log of the likelihood ratio) is often difficult to derive, [79] showed that, just like the Pearson Chi-Squared test, the Likelihood Ratio test statistic also

is 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 are usually going to feature 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. Despite solving the problem of data being needlessly thrown away, this approach still shares the problem of an improper nominal type I error rate and inaccurate p-values. The bright side to all this, as is about to be shown via simulations, is that both the Chi-Square test and the Likelihood Ratio Test tend to be on the overly conservative side, so it seems like we can trust results denoting a rejection of the null even more than usual, at the expense of elevated Type II errors.

To avoid direct reliance on theoretical test distributions, [11] 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 from the parameter estimates, looking at the distribution of test statistics and checking whether the raw test statistic falls outside this range. While in [11] they 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.

3.3.1 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 [69] to generate the probabilities of the true sets of preferences. 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 calculated 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 realistic sets of parameters.

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 between 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. After that, each subject's true probabil-

ities 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. 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 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 .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 amount of data one would collect with the individual model and 24 representing a relatively large amount 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.

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

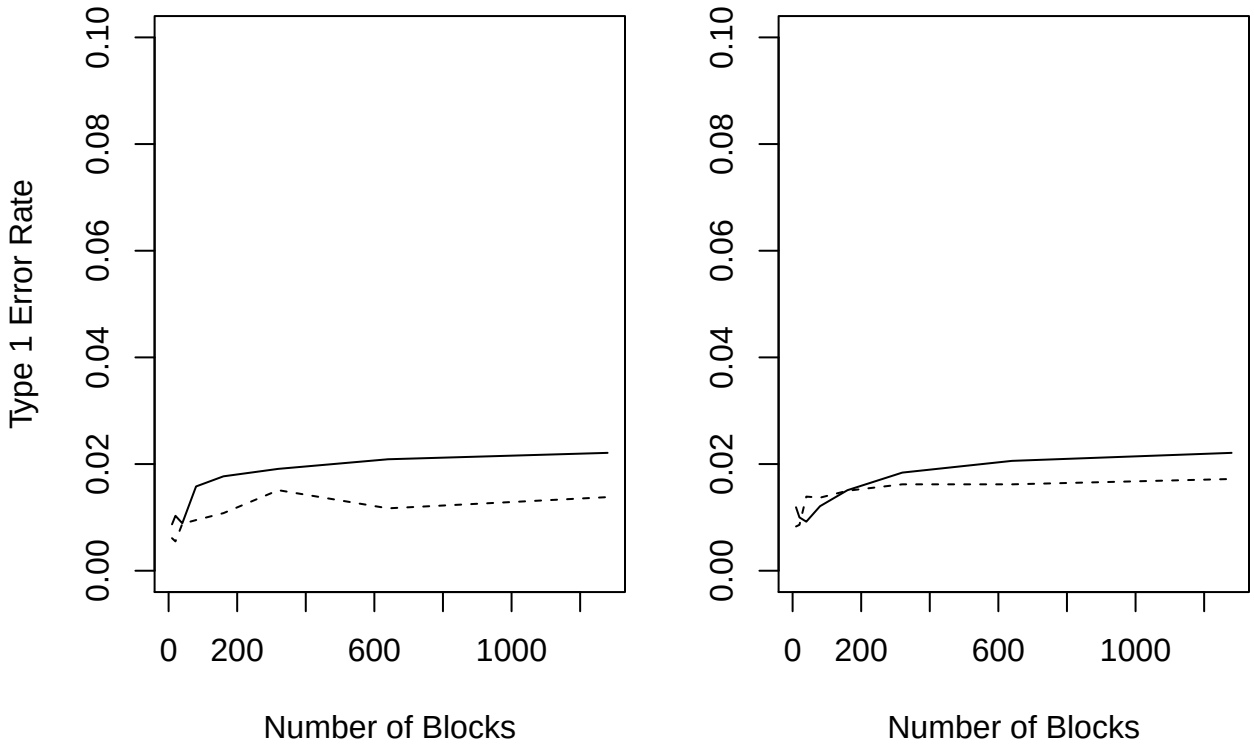

Figure 3.1: Type 1 Error Rate vs. Number of Blocks for the Likelihood Ratio Test (dashed) and Chi Square (solid). On the left are results via simulation with the probit parameter generation approach, and on the right the dirichlet approach. (10000 simulations)

case of bootstrapping if the 95% confidence interval of either of the two possible intransitive probabilities did not include any value below .001. For the monte-carlo procedure, the null was said to be rejected if 95% or more of the simulated LRT statistic distribution came out below the real LRT statistic. 2000 experiments were simulated for each condition utilizing the University of California, Irvine's high performance computing cluster.

3.3.2 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 3.2both using the full data (on the left) and data reduced as suggested in [10]. As we can clearly see, utilization of the full data provides a substantial reduction in MSE, yielding more accurate estimates with the same

Table 3.1: Power and Level for each frequentist hypothesis testing method for both parameter generation approaches.

	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

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 as can be seen in figure 3.1. What we can see from these simulations is 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 3.1. 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. We can see that while still being overly conserative, the Bootstrap procedure yields by far the highest power. Although Monte-Carlo ends up getting the type I error rate pretty close to the nominal alpha level, we can see that the power is actually worse relative to the other methods. This is not too surprising, because in a sense the Monte-Carlo simulations simulate 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.

3.4 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 utilizing group-level information. Hierarchical models are powerful tools that have been proven to provide more accurate parameter estimates than non-Hierarchical Bayesian models, as measured by MSE [19]. Bayesian statistics in general differs in that rather than providing point estimates for parameters, posterior distributions are extracted from the data according to the model specification.

Although it was applied to a different variation of the True and Error model involving only two preferences per set instead of three, [33] had 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 utilized 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:

$$
\begin{aligned}
X_{i,j} &\sim Normal(\mu_j, \tau_j) \\
p_{i,j} &= \frac{e^{X_{i,j}}}{\sum_{j=1}^{8} e^{X_{i,j}}} \\
\mu_j &\sim Normal(0, 1) \\
\tau_j &\sim Exponential(1)
\end{aligned}
$$

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 here 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 and then transformed into probabilities according to a softmax function taking in the rest of subject i's Xs.

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:

$$2e_{i,k} \sim Beta(1, 2) \tag{3.6}$$

It should be noted that in [**?**] implementation, these were drawn from uniform distributions. The non-hierarchical version used in this paper differs from the hierarchical model only in that the ps are instead drawn from a flat dirichlet distribution.

To implement this in JAGS it is possible to calculate the probability of all 64 possible combinations of preference orderings beforehand and treat the combination as coming from a categorical distribution similar to the approach found in [33]. Alternatively one can upload

Table 3.2: 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

a vector with a one for each observed combination as data and treat the observation as coming from a Bernoulli distribution with the corresponding probability as in the code given in the appendix.

MSE performance for estimating the ps in each model from 100 simulations of each case is shown in Table 3.2, both in posterior distribution as well as in point estimation (in this case the mean of the posterior distribution). We see here 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 performs better despite the fact that the simulation approaches do not generate people's parameters according to softmax transformed normally distributed variables as specified by the model, highlighting its robustness.

3.4.1 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 distribution of the parameters for comparing two models [30]. In this case, one can practically calculate the Bayes factor using a spike-and-slab approach by adding a Bernoulli distributed indicator parameter for transitivity with prior probability of 1/2. When the indicator for transitivity 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 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 Proportion of the time the indicator shows intransitivity divided by the proportion of the time the indicator indicates transitivity 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 .5 chance of being truly transitive or intransitive. Results of this can be seen in Table 4.2. While some researchers might want to avoid having formal cutoffs for Bayes Factors, proportion of Bayes Factors greater than 1, 3 and 10 were reported along with corresponding type 1 error rates. [76] 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

Table 3.3: 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

a type-1 error rate of .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 .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.

3.5 Discussion

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 utilizing all the data, because the proper null distributions are still unknown and their explored frequentist

tests tend to be overly conservative. Beyond that, Bayesian methods have the advantage of natively providing a framework with easily interpretable confidence intervals for estimated parameters. While bootstrapping and utilizing all of the data using a Likelihood Ratio Test instead of the Chi-Squared test on a reduced number of degrees of freedom gives a substantial gain in performance, it still fails to match the advantages of the Bayesian approach.

3.6 Appendix: Hierarchical Model JAGS Code:

```
model{

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

  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]))
}
```

```
}
```

Chapter 4

Transitivity Violations in Probabilistic and Delay Discounting

4.1 Introduction

In many cases people opt for smaller monetary rewards given sooner over larger ones awarded later. Similarly, people often choose smaller but more probable rewards over larger, less probable rewards. A natural question that arises for behavioral economists and psychologists to answer is how much less is a monetary reward worth given the wait time to receive it, or given the probability that the reward won't be received at all. Implicit in this question is the assumption of transitivity of preference, one of the fundamental axioms of rational choice theory, which states that if any option A is preferred over any option B, and option B is prefferred over any option C, then C cannot be preferred over option A. The present study explores whether this assumption really holds.

It has been widely suggested that there are inherent similarities in the domains of probabilistic and temporal discounting. (e.g. [22, 40, 44, 17, 48]). Although seemingly unrelated,

previous research has suggested that both discounting curves seem to be successfully modeled with the same hyperbolic function, which in both cases can be derived under the assumption that the subjective value is proportional to the expected reward over time [43]. In this case, the functional form of the discounting curve describing depreciation in value is equivalent when odds against winning is substituted for time. Although our methods to study the question of transitivity in these domains does not rely on any assumptions of functional form, or even existence of such a discounting curve, it is useful to know it to really understand the previous research in the matter.

$$V_{\$} = \frac{A}{1 + k_i * D} \tag{4.1}$$

Describes the subjective monetary value for person i with delay discounting rate k_i at amount A. In other words, when a delayed option has a subjective monetary value of $V_{\$}$, an immediate monetary option greater than $V_{\$}$ is more likely to be preferred, while an immediate option less than $V_{\$}$ is more likely to be rejected in favor of the delayed reward. Higher values of k_i indicate that subject i deems monetary rewards to decrease in value more rapidly with respect to the magnitude of the delay.

Similarly:

$$V_{\$} = \frac{A}{1 + h_i * \theta} \tag{4.2}$$

Describes the subjective monetary value for person i with probabilistic discounting rate h_i at amount A, where $\theta = \frac{1-p}{p}$ is the odds against success.

There has been some evidence that delay discounting may actually be sensitive to context,

and thus has the potential to induce violations of transitivity. [23] investigated this and concluded that when subjects compare two delayed rewards, the tendency was for the subjects to discount the sooner reward less than they would if they were comparing it to the immediate reward. The analysis by [23], however, was conducted using a parametric test assuming a particular functional form instead of investigating transitivity directly with no such assumptions as is done in this study. They named this context effect the Common Aspect Attenuation Hypothesis (CAAH), and referred to the rejected null result as the Present Value Comparison Hypothesis (PVCH). Under the CAAH as they formalized it, the way two delayed options get discounted is slightly different than an immediate option with a delay. Specifically, drawing from the proposed functional form from [23] the sooner option gets discounted according to:

$$V_{\$,s} = \frac{A_s}{1 + w_{k,i} k_i * D_s} \tag{4.3}$$

while the later option gets discounted according to:

$$V_{\$,l} = \frac{A_l}{1 + w_{k,i} k_i * D_s + k_i * (D_l - D_s)} \tag{4.4}$$

It should be noted that when w=1, or when D or θ are zero for one of the options, the CAAH reduces to the PVCH.

There hasn't been a parallel experiment done for probabilistic discounting, but given that the common functional form for the two domains is the same, it would seem natural to extend a possible CAAH to probabilistic discounting as follows:

$$V_{\$,m} = \frac{A_m}{1 + w_{h,i} h_i * \theta_m} \tag{4.5}$$

For the more probable option, while the less probable option's subjective monetary value

would be:

$$V_{\$,l} = \frac{A_l}{1 + w_{h,i} h_i * \theta_s + h_i * (\theta_l - \theta_m)} \tag{4.6}$$

Although probabilistic and temporal discounting share the same form, there have also been findings that have set them apart as well. For example, it has been found that while in temporal discounting the rate of discounting with respect to time gets smaller with larger monetary amounts, the opposite is true for the rate of probabilistic discounting with respect to odds against winning [25, 24]. Because there are such differences, it's not implausible that probabilistic discounting would not yield results characterized by the same formalism as the CAAH in delay discounting. For an in-depth review of probabilistic and temporal discounting and their relations to each other, see [22].

The CAAH gives us reason to believe that we may find violations of transitivity in delay discounting and probabilistic discounting if subjects are asked to compare two delayed or two probabilistic options with each other, but what about choosing between a delayed option vs a probabilistic one? When domains cross, does subjective value get preserved, or are there violations of transitivity? There hasn't been much investigation into this, but results from [40] hint that there might be intransitivities in such cases too via a parametric analysis based on the hyperbolic discounting function.

The approach we take is unlike that in [23] and [40] in that there is no reliance on the functional form of discounting functions. Instead, we've systematically selected sets of triples and utilized a repeated measures approach. We use a True and Error (TE) model [10] to analyze the data, which allows us to study concurrent sets of preferences instead of resorting to stochastic transitivity and triangle inequalities like in [54] which rely on assumptions of IID, often known to be faulty in repeated measures designs [9].

4.2 Stimulus Generation and Experimental Procedure

A previous, unpublished study on probabilistic and delay discounting with 30 participants was used to find optimal sets of stimuli (see attached supplementary material). Data in this study was analyzed using a hierarchical Bayesian modeling framework based on the hyperbolic discounting model and the functional form for the CAAH proposed by [23]. Although the CAAH form was designed for delay discounting specifically, the same form was adopted also for probabilistic discounting. The study also included questions asking subjects to pick between a probabilistic and delayed option. This was modeled with the use of an additive term representing biased preference toward probabilistic or temporally discounted options under the context of a comparison between the two. The results of the analysis supported the CAAH in both the delay and probabilistic domains and suggested there may be violations of transitivity when you mix the two domains. A detailed writeup of this study is available in the attached supplementary material.

Starting with pre-picked immediate/certain options, a brute-force grid-search simulation approach using the individual parameter estimates of the 30 participants was utilized to find 4 sets of triples in each of the the 3 domains that would yield the highest number of people committing violations of transitivity. Monetary amounts were allowed to reach up to $4000, delays up to 48 months, and probabilities as low as 0.2. The selected sets of triples are shown in Table 4.1.

In anticipation of needing to use the True and Error model, all pairs among the 12 sets of triples were each presented twice per block, one time in reverse order. Filler paired comparison questions generated in the same way as the previous study were thrown in between the pairs corresponding to the selected sets of triples. Subjects were mostly brought in for two separate 80 minute sessions with questions automatically generated until the timer finished, but in some cases accomodations were made so that one or more of the

Table 4.1: Triples used

Triple	A	B	C
Delay			
1	\$290 right now	\$568 in 28 months	\$661 in 34 months
2	\$330 right now	\$972 in 18 months	\$2716 in 44 months
3	\$440 right now	\$796 in 22 months	\$1152 in 48 months
4	\$550 right now	\$1758 in 36 months	\$2275 in 46 months
Probabilistic			
1	\$310 for certain	\$771 with a chance of 0.56	\$3539 with a chance of 0.29
2	\$340 for certain	\$1164 with a chance of 0.4	\$2719 with a chance of 0.26
3	\$430 for certain	\$1144 with a chance of 0.3	\$1769 with a chance of 0.24
4	\$550 for certain	\$1758 with a chance of 0.5	\$2706 with a chance of 0.45
Mixed			
1	\$280 right now	\$2363 with a chance of 0.38	\$1619 in 39.5 months
2	\$315 right now	\$536 with a chance of 0.27	\$1273 in 22 months
3	\$532 right now	\$3792 with a chance of 0.31	\$3861 in 19.5 months
4	\$827 right now	\$3937 with a chance of 0.22	\$2858 in 8.5 months

sessions would be slightly shorter. In total there were 27 participants, mostly Psychology undergraduates with some Graduate Students from Cognitive Science participating as well.

The responses on the trials from the old experiment were fit to the old model, and a posterior predictive check was performed to see if the predictions from the old model matched the observations on the repeated measures. The old model performed poorly on these checks, and in fact severely underestimated the true number of transitivity violations that were observed in the probabilistic and mixed domains. Although this approach delivered strong results of transitivity violations as was desired, it seems like the old model needs revision to capture people's true behavior in these domains.

4.3 Analysis Methods

We used the procedure recommended in [8] to check for evidence of violations of independence assumptions inherent to many existing models of transitivity. To ensure satisfactory power all questions in all domains were analyzed. To properly adhere to the procedure, it was only possible to use the number of repetitions of the least repeated question for each person. 10000 samples were run, simulating random permutations of the question order. We found that 20 out of the 27 participants had evidence of violating independence at the $\alpha = .05$ level, with one more participant with a p-value just .0012 away from significance, and the rest failing to reject the null hypothesis of response independence. Because of this evidence that at least for most people the response independence assumption was not supported, we opted to use the True and Error model.

The previous chapter of this dissertation offered substantial improvements to True and Error analysis in accuracy and data efficiency. Although several new methods of analysis were introduced, the one with most improvements and suitability for this experimental task is the Hierarchical Bayesian implementation of the True and Error model, so that is what was utilized in the present study.

4.4 Results

Table 4.2 summarizes the results of the tests for violations of transitivity. The columns show the proportion of individuals whose Bayes Factors favored violations of transitivity crossing thresholds of 1, 3, and 10 respectively. Although Bayes Factors are fundamentally different in nature from p-values, to give a sense of the strength of the evidence the previous dissertation chapter showed through simulation that a Bayes Factor threshold of 1 corresponded to a type 1 error rate (α) of roughly .08 for the model used, 3 corresponded to an α level of

Table 4.2: Proportion whose Bayes Factors favored the Intransitive Model (N=27)

Triple	BF>1	BF>3	BF>10
Delay			
1	0.22	0.00	0.00
2	0.11	0.00	0.00
3	0.22	0.15	0.04
4	0.15	0.00	0.00
Highest	0.52	0.15	0.04
Probabilistic			
1	0.48	0.44	0.30
2	0.30	0.19	0.15
3	0.37	0.30	0.22
4	0.56	0.56	0.52
Highest	0.67	0.59	0.56
Mixed			
1	0.44	0.37	0.37
2	0.44	0.37	0.37
3	0.63	0.56	0.48
4	0.33	0.30	0.30
Highest	0.81	0.78	0.70
All			
Highest	0.96	0.89	0.81

roughly .03-.04 , and a threshold of 10 corresponded to an α of roughly .01. The rows in 4.2 labeled "Highest" correspond to the proportion of people whose highest Bayes Factor in the respective domain exceeded each threshold. Tables containing raw proportions of transitivity violations in either direction are included in the appendix, along with model estimates of the true probabilities of each direction of transitivity violations.

Although the CAAH was initially conceived as a phenomenon of delay discounting, we actually find that the delay domain was the only one out of the three that was mostly void of violations of transitivity. The only delay triple that showed any signs of violations of transitivity was the third one, with only 4 out of 27 people exceeding the BF>3 threshold and just 1 exceeding the BF>10 threshold. Examining the tables in the appendix, the cases of transitivity violations detected were actually in the opposite direction than that predicted by the CAAH. Since the prevalence of this detection was so low, it's possible that this result could be a mere fluke, and not reflective of a larger pattern. It should be stressed that although these results didn't support the CAAH, they are not enough to dismiss the CAAH outright. Under the model for the CAAH from [23], the proportion of triples any one individual would be intransitive with is quite small, so it's plausible that the ones selected simply weren't the right ones to detect a violation of transitivity for any of the participants.

Far more violations of transitivity were detected in the Probabilistic domain, and every one of them was in the direction that would be predicted by the probabilistic discounting analogue to the CAAH. Although the direction of intransitivity is consistent with that predicted by the CAAH, the results observed here were actually too strong for the functional form of the CAAH analogue to hold. There is no set of parameters under that model that would allow for an individual subject to have violations of transitivity over all four probabilistic triples, and yet there were multiple people where this was observed decisively. In many instances, people responded intransitively over 90% of the time. Overall, 59% of people had at least one Bayes Factor favoring a transitivity violation in at least on case past the BF>3 threshold,

and 56% of people had at least one past the BF>10 threshold. The 4th triple was the one that people were most intransitive with, with 56% crossing the BF>3 threshold and 52% crossing the BF>10 threshold, meaning that all but one person who showed intransitivity in the Probabilistic domain were intransitive with that triple, and many were also intransitive with the others.

The mixed domain ended up in total having the most violations of transitivity. 78% of people crossed BF>3, and 70% crossed BF>10 in for at least one of the four triples. Results in this case were very strong, with multiple cases of people choosing intransitively 100% of the time. The third triple ended up inducing the most violations of transitivity, and interestingly was the only case were transitivity violations were detected in both directions. All other triples had the probabilistic option chosen over the delayed option in cases of transitivity violations, but in the case of the third triple chosing the delayed option over the probabilistic was more prominent under transitivity violations. With minor exceptions, most of those with transitivity violations favoring delayed above probabilistic options in the third triple did not exhibit transitivity violations for the other triples in the mixed domain.

4.5 Discussion

Overall the results were stronger than anticipated in all but the delay domain. It seems like there's a need to rethink the exact mathematical formalism behind these strong context effects in probabilistic discounting, because simply transferring over the model of the CAAH from [23] into the probabilistic domain doesn't seem to cut it. The transitivity violations in the mixed domain were strongest and most common, but the model from the supplementary paper was insufficient in capturing everything that was going on. It's clear that this needs to be explored more in depth if an accurate model that accounts for context is to be developed.

4.6 Appendix

Table 4.3: Proportion of time $A > B > C > A$ (Raw Data)

	Delay				Probabilistic				Mixed			
ID	1	2	3	4	1	2	3	4	1	2	3	4
1	0.23	0.03	0.00	0.04	0.87	0.70	0.86	0.93	0.73	0.65	0.13	0.57
2	0.08	0.14	0.06	0.08	0.00	0.00	0.03	0.03	0.05	0.05	0.00	0.05
3	0.00	0.05	0.08	0.11	0.35	0.10	0.42	0.04	0.14	0.15	0.04	0.11
4	0.09	0.09	0.05	0.05	0.55	0.11	0.22	0.23	0.68	0.65	0.20	0.75
5	0.14	0.00	0.07	0.06	0.58	0.17	0.58	0.17	0.07	0.19	0.00	0.25
6	0.04	0.00	0.08	0.07	0.81	0.32	0.71	0.43	0.07	0.04	0.00	0.07
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.70	1.00	0.00	0.00
8	0.09	0.03	0.03	0.00	0.03	0.07	0.00	0.03	0.12	0.10	0.00	0.03
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.87	0.47	0.60
10	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.91	0.86	0.23	0.68
11	0.06	0.00	0.00	0.00	0.28	0.28	0.22	0.78	0.00	0.00	0.00	0.00
12	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.25	0.00	0.04
13	0.00	0.00	0.00	0.00	0.32	0.14	0.59	0.91	0.05	0.05	0.00	0.00
14	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.06	0.94	0.97	0.56	0.61
15	0.00	0.05	0.00	0.00	0.23	0.41	0.36	0.50	0.14	0.82	0.00	0.18
16	0.06	0.00	0.06	0.06	0.94	0.07	0.06	0.44	0.06	0.25	0.00	0.06
17	0.00	0.06	0.05	0.11	0.61	0.44	0.62	0.45	0.25	0.33	0.06	0.18
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.18	1.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	1.00	0.00	1.00
20	0.00	0.00	0.00	0.00	0.80	0.10	0.20	0.70	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.12	0.75	0.00	0.12	0.75	0.25	0.25	0.00	0.00
22	0.00	0.33	0.00	0.17	0.50	0.25	0.88	0.67	0.00	0.17	0.00	0.00
23	0.00	0.00	0.00	0.00	0.95	0.72	0.94	0.95	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	0.12	0.00	0.19	0.00	1.00	1.00	0.88	0.88
25	0.10	0.08	0.25	0.14	0.93	0.20	0.75	0.50	0.00	0.14	0.07	0.00
26	0.09	0.17	0.00	0.00	0.54	0.58	0.59	0.46	0.04	0.17	0.00	0.18
27	0.10	0.10	0.30	0.06	0.90	0.20	0.44	0.85	0.15	0.30	0.00	0.05

Table 4.4: Proportion of time $A < B < C < A$ (Raw Data)

	Delay				Probabilistic				Mixed			
ID	1	2	3	4	1	2	3	4	1	2	3	4
1	0.00	0.03	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
3	0.05	0.15	0.04	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.18	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00
5	0.07	0.14	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.00
6	0.21	0.11	0.19	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00
7	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.16	0.00	0.30	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
9	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.09	0.09	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.00
12	0.00	0.05	0.09	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.27	0.04	0.50	0.05	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
14	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.05	0.23	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00
16	0.19	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00
17	0.06	0.06	0.10	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00
21	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00
22	0.00	0.33	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00
23	0.00	0.00	0.06	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.00
24	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00
26	0.05	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00
27	0.00	0.05	0.05	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.00

Table 4.5: Proportion of time $A > B > C > A$ (True Probability Estimate)

	Delay				Probabilistic				Mixed			
ID	1	2	3	4	1	2	3	4	1	2	3	4
1	0.02	0.00	0.01	0.01	0.90	0.97	0.98	0.98	0.73	0.70	0.04	0.64
2	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.02	0.04	0.01	0.01	0.03	0.00	0.12	0.02	0.03	0.03	0.02	0.00
4	0.00	0.02	0.00	0.01	0.67	0.04	0.01	0.26	0.73	0.86	0.03	0.85
5	0.05	0.03	0.04	0.03	0.74	0.01	0.65	0.17	0.03	0.03	0.01	0.09
6	0.00	0.01	0.05	0.03	0.99	0.03	0.97	0.73	0.01	0.00	0.01	0.01
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.93	0.97	0.02	0.01
8	0.01	0.00	0.02	0.01	0.00	0.05	0.00	0.01	0.01	0.02	0.01	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99	0.97	0.22	0.85
10	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.99	0.96	0.13	0.77
11	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.96	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.03	0.01	0.00	0.00
13	0.00	0.00	0.02	0.01	0.03	0.00	0.78	0.93	0.00	0.00	0.00	0.01
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99	0.99	0.50	0.93
15	0.00	0.00	0.00	0.00	0.00	0.71	0.03	0.77	0.05	0.98	0.01	0.10
16	0.03	0.04	0.01	0.02	0.99	0.01	0.00	0.82	0.01	0.03	0.01	0.00
17	0.02	0.04	0.03	0.02	0.84	0.46	0.85	0.62	0.32	0.41	0.02	0.05
18	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.99	0.98	0.04	0.98
19	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.99	0.98	0.01	0.98
20	0.01	0.01	0.00	0.01	0.97	0.05	0.01	0.89	0.01	0.01	0.01	0.01
21	0.00	0.00	0.00	0.01	0.88	0.01	0.01	0.81	0.23	0.04	0.02	0.01
22	0.00	0.05	0.03	0.02	0.44	0.03	0.96	0.89	0.04	0.06	0.02	0.02
23	0.03	0.00	0.01	0.00	0.99	0.93	0.99	0.97	0.00	0.00	0.00	0.01
24	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.98	0.98	0.88	0.97
25	0.02	0.01	0.01	0.01	0.99	0.02	0.96	0.75	0.00	0.01	0.03	0.01
26	0.05	0.03	0.02	0.01	0.60	0.82	0.63	0.76	0.01	0.02	0.03	0.15
27	0.07	0.01	0.00	0.02	0.99	0.02	0.17	0.97	0.10	0.06	0.01	0.01

Table 4.6: Proportion of time $A < B < C < A$ (True Probability Estimate)

	Delay				Probabilistic				Mixed			
ID	1	2	3	4	1	2	3	4	1	2	3	4
1	0.01	0.01	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.26	0.01
2	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.03	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
4	0.00	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.27	0.00
5	0.05	0.04	0.16	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.72	0.00
6	0.00	0.02	0.30	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.91	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00
8	0.02	0.00	0.45	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
10	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
11	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
12	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
13	0.01	0.00	0.78	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.90	0.00
16	0.05	0.03	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.00
17	0.03	0.04	0.06	0.03	0.00	0.00	0.00	0.01	0.00	0.01	0.65	0.00
18	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
19	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
20	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.00
21	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.02	0.00	0.23	0.00
22	0.00	0.06	0.08	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.86	0.01
23	0.02	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
25	0.04	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.87	0.00
26	0.04	0.02	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.00
27	0.04	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00

Bibliography

[1] G. E. Alexander. *A Mathematical Explication of Human Psychology*. PhD thesis, 2017.

[2] R. Anders and W. H. Batchelder. Cultural consensus theory for the ordinal data case. *Psychometrika*, 80(1):151–181, 2015.

[3] R. Anders, Z. Oravecz, and W. Batchelder. Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, 61:1–13, 2014.

[4] W. H. Batchelder and R. Anders. Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, 56(5):316–332, 2012.

[5] W. H. Batchelder, R. Anders, and Z. Oravecz. Cultural consensus theory. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 5:201–264, 2018.

[6] W. H. Batchelder, A. Strashny, and A. K. Romney. Cultural consensus theory: Aggregating continuous responses in a finite interval. In *International Conference on Social Computing, Behavioral Modeling, and Prediction*, pages 98–107. Springer, 2010.

[7] D. J. Bem. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100:407–425, 2011.

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

[9] M. H. Birnbaum. Testing mixture models of transitive preference: Comment on regenwetter, dana, and davis-stober (2011). 2011.

[10] M. H. Birnbaum. True-and-error models violate independence and yet they are testable. *Judgment and Decision making*, 8(6):717, 2013.

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

[12] M. H. Birnbaum and E. G. Quispe-Torreblanca. Temap2. r: True and error model analysis program in r. *Judgment and Decision Making*, 13(5):428, 2018.

[13] G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[14] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Bettencourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 2017.

[15] R. S. Chhikara and J. L. Folks. *The inverse Gaussian distribution: Theory, methodology, and applications*. Marcel Dekkar, New York, 1989.

[16] J. Correll, B. Park, C. Judd, and C. Wittenbrink. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6):1314–1329, 2002.

[17] W. Du, L. Green, and J. Myerson. Cross-cultural comparisons of discounting delayed and probabilistic rewards. *The Psychological Record*, 52(4):479–492, 2002.

[18] E. N. Dzhafarov. The structure of simple reaction time to step-function signals. *Journal of Mathematical Psychology*, 36:235–268, 1992.

[19] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[20] A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

[21] P. Gomez, R. Ratcliff, and M. Perea. The overlap model: A model of letter position coding. *Psychological Review*, 115(3):577, 2008.

[22] L. Green and J. Myerson. A discounting framework for choice with delayed and probabilistic rewards. *Psychological bulletin*, 130(5):769, 2004.

[23] L. Green, J. Myerson, and E. W. Macaux. Temporal discounting when the choice is between two delayed rewards. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):1121, 2005.

[24] L. Green, J. Myerson, and E. McFadden. Rate of temporal discounting decreases with amount of reward. *Memory & cognition*, 25(5):715–723, 1997.

[25] L. Green, J. Myerson, and P. Ostaszewski. Amount of reward has opposite effects on the discounting of delayed and probabilistic outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(2):418, 1999.

[26] J. M. Haaf and J. N. Rouder. Developing constraint in Bayesian mixed models. *Psychological Methods*, 22(4):779–798, 2017.

[27] J. M. Haaf and J. N. Rouder. Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin and Review*, 2019.

[28] E. Harmon-Jones and P. A. Gable. Neural activity underlying the effect of approach-motivated positive affect on narrowed attention. *Psychological Science*, 20(4):406–409, 2009.

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

[30] H. Jeffreys. Theory of probability, clarendon, 1961.

[31] J. K. Kruschke. *Doing Bayesian data analysis, 2nd edition: A tutorial with R, JAGS, and Stan*. Academic Press, Waltham, MA, 2014.

[32] D. R. J. Laming. *Information Theory of Choice-Reaction Times*. Academic Press, London, 1968.

[33] M. D. Lee. Bayesian methods for analyzing true-and-error models. *Judgment and Decision making*, 13(6):622, 2018.

[34] M. D. Lee, M. Steyvers, M. De Young, and B. Miller. Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, 4(1):151–163, 2012.

[35] M. D. Lee and E.-J. Wagenmakers. *Bayesian cognitive modeling: A practical course*. Cambridge University Press, 2013.

[36] J. Levav and G. J. Fitzsimons. When questions change behavior: The role of ease of representation. *Psychological Science*, 17(3):207–213, 2006.

[37] S. Lichtenstein and P. Slovic. Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, 89(1):46, 1971.

[38] S. Lo and S. Andrews. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6:1171, 2015.

[39] R. D. Luce. *Response times*. Oxford University Press, New York, 1986.

[40] A. Luckman, C. Donkin, and B. R. Newell. Can a single model account for both risky choices and inter-temporal choices? testing the assumptions underlying models of risky inter-temporal choice. *Psychonomic bulletin & review*, pages 1–8, 2017.

[41] S. A. Massar, J. Lim, K. Sasmita, and M. W. Chee. Rewards boost sustained attention through higher effort: A value-based decision making approach. *Biological psychology*, 120:21–27, 2016.

[42] J. R. Michael, W. R. Schucany, and R. W. Haas. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

[43] J. Myerson and L. Green. Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, 64(3):263–276, 1995.

[44] J. Myerson, L. Green, J. S. Hanson, D. D. Holt, and S. J. Estle. Discounting delayed and probabilistic rewards: Processes and traits. *Journal of Economic Psychology*, 24(5):619–635, 2003.

[45] J. Neyman and E. S. Pearson. 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, 1933.

[46] K. Pearson. 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, 1900.

[47] M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.

[48] H. Rachlin, A. Raineri, and D. Cross. Subjective probability and delay. *Journal of the experimental analysis of behavior*, 55(2):233–244, 1991.

[49] R. Ratcliff. A theory of memory retrieval. *Psychological Review*, 85:59–108, 1978.

[50] R. Ratcliff. Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, 86:446–461, 1979.

[51] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114:510–532, 1993.

[52] R. Ratcliff, A. Thapar, and G. McKoon. The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, 16:323–341, 2001.

[53] B. Reddi and R. Carpenter. Accuracy, information and response time in a saccadic decision task. *Journal of Neurophysiology*, 90:3538–3546, 2003.

[54] M. Regenwetter, J. Dana, and C. P. Davis-Stober. Transitivity of preferences. *Psychological review*, 118(1):42, 2011.

[55] A. K. Romney, S. C. Weller, and W. H. Batchelder. Culture as consensus: A theory of culture and informant accuracy. *American anthropologist*, 88(2):313–338, 1986.

[56] J. N. Rouder. Are unshifted distributional models appropriate for response time? *Psychometrika*, 70:377–381, 2005.

[57] J. N. Rouder and J. M. Haaf. A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, in press.

[58] J. N. Rouder, M. S. Pratte, and R. D. Morey. Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin and Review*, 17:427–435, 2010.

[59] J. N. Rouder, J. M. Province, R. D. Morey, P. Gomez, and A. Heathcote. The lognormal race: a cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80:491–513, 2015.

[60] J. N. Rouder, F. Tuerlinckx, P. L. Speckman, J. Lu, and P. Gomez. A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, 15(1201-1208), 2008.

[61] W. Schwarz. The ex-Wald distribution as a descriptive model of response times. *Behavioral Research Methods, Instruments, and Computers*, 33:457–469, 2001.

[62] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[63] J. R. Simon. Reactions toward the source of stimulation. *Journal of Experimental Psychology*, 81:174–176, 1969.

[64] S. Sternberg. High-speed scanning in human memory. *Science*, 153:652–654, 1966.

[65] M. Stone. Models for choice-reation time. *Psychometrika*, 25:251–260, 1960.

[66] J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–662, 1935.

[67] J. E. Thiele, J. M. Haaf, and J. N. Rouder. Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology*, 81:40–54, 2017. Revision submitted 3/17.

[68] L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[69] L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[70] A. Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

[71] R. Ulrich and J. O. Miller. Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, 37:513–525, 1993.

[72] M. Usher and J. L. McClelland. On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, 108:550–592, 2001.

[73] T. Van Zandt. How to fit a response time distribution. *Psychonomic Bulletin and Review*, 7:424–465, 2000.

[74] D. Vickers and P. Smith. Accumulator and random-walk models of psychophysical discrimination: A counter-evaluation. *Perception*, 14:471–497, 1985.

[75] E. J. Wagenmakers and S. Brown. On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, 114:830–841, 2007.

[76] E.-J. Wagenmakers, R. D. Morey, and M. D. Lee. Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3):169–176, 2016.

[77] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[78] R. R. Wilcox. How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53(3):300, 1998.

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

Appendix A

Supplementary Material

Attached is a write-up of the experiment that preceded that which was described in the fourth chapter.

Probabilistic and Temporal Discounting: A Thurstonian Approach

Pele Schramm

University of California, Irvine

Abstract

This study further explores the mechanics of both probabilistic and temporal discounting and their relation to each other under a hierarchical Bayesian modeling framework that includes a stochastic model for the raw data. Strong evidence for context effects were found at the individual level for forced-choice comparisons between two delayed monetary rewards, two probabilistic monetary rewards, and comparisons between a delayed and a probabilistic reward. Findings from previous experiments regarding changes in discounting rate with respect to monetary amount in both delay and probabilistic domains were confirmed true in most individuals.

Keywords:
Probabilistic, Temporal, Discounting, Thurstone, Hierarchical, Bayesian, Decision Making

1. Introduction

The subjective devaluation of a monetary reward given the time it will take to receive the reward is a classic area of study within behavioral economics and decision science, as is the devaluation of a potential monetary reward with respect to the probability of receiving it. Many researchers have suggested that there is an inherent similarity between these two (e.g. Green and Myerson (2004); Luckman et al. (2017); Myerson et al. (2003); Du et al. (2002); Rachlin et al. (1991)). Both probabilistic and temporal discounting seem to be well represented under a hyperbolic discounting model, which itself can be derived under the assumption that the subjective value is proportional to the expected reward over time (Myerson and Green, 1995). In this case, the functional form of the discounting curve describing depreciation in value is equivalent when odds against winning is substituted for time.

There have been a number of interesting findings regarding both probabilistic and temporal discounting. Firstly, it has been found that while in temporal discounting the rate of discounting with respect to time gets smaller with larger monetary amounts, the opposite is true for the rate of probabilistic discounting with respect to odds against winning (Green et al., 1999, 1997). Green et al. (2005) found that when comparing two delayed rewards, subjects tend to discount the sooner reward less than they would if they were comparing it to an immediate reward. This effect is referred to as the Common Aspect Attenuation Hypothesis (CAAH), vs the Present Value Comparison Hypothesis (PVCH) which states that this context effect isn't there. Under a discounting framework, there doesn't seem to have been much investigation of this effect under the probabilistic counterpart. There also has not been much investigation into subjects comparing probabilistic vs temporal rewards, although there has been some evidence that there is a potential context effect there (Luckman et al., 2017). Besides research that has delved into the mechanics of probabilistic and temporally discounted decision making, there has also been a large body of research that has found connections between temporal discounting rates and various addictions such as heroin addiction (Kirby et al., 1999), alcoholism (Petry, 2001), cigarette addiction (Baker et al., 2003), as well as other behavioral tendencies. For an in-depth review, see Green and Myerson (2004).

Most of the previous research has used a psychophysical procedure developed by Rachlin et al. (1991) to measure discounting curves. In this procedure, the discounting curve for a single monetary amount is measured by iteratively asking subjects whether they would prefer that monetary amount at a certain delay or a smaller monetary amount immediately, each time adjusting the immediate amount higher or lower until there is a preference reversal. This is then repeated at a number of different delay points, plotting the points of preference reversal and using them to fit the discounting curve. While this procedure does get at the question at hand, there are a number of drawbacks. Firstly, there is susceptibility to order effects and memory effects. These can be diminished through varying the starting position of the immediate option to start high or low, which works well if finding an aggregate function, but may still suffer from memory effects when evaluating individual discounting functions. Something else that is lacking is a theory on the probability of choosing one option over the other, creating a level of separation between the statistical analysis and the raw data and inhibiting ability to make probabilistic predictions of responses to new pairs of options using the data. To study context effects, this procedure must be combined with specific other items, and with no probabilistic framework for individual responses it becomes difficult to effectively study context

effects at an individual level rather than aggregate to really determine what proportion, if any at all, really have susceptibility to context effects and how.

This study aims to resolve these issues, allowing us to efficiently study a variety of context effects in discounting for a range of different monetary values in a manner that effectively differentiates between individuals and aggregate. In addition to looking at the CAAH at the individual level in both probabilistic and temporal domains, this study also investigates whether individual subjects are biased more towards favoring delayed options or probabilistic options when given a choice between the two.

A useful framework for formalizing subjective value as it relates to choice probabilities under paired comparisons can be taken from Thurstone (1927). The Thurstonian model assumes that the subjective value of each option in a paired comparison task is estimated at each comparison in a Gaussian distributed fashion, and the option which at that comparison was estimated to have a higher subjective value is chosen. The mean of this Gaussian distribution can be treated as the true subjective value, while the variance corresponds to the degree to which the subjective value of the option may vary. It is important to note that in this context, subjective value is measured on a latent scale, which is not necessarily linearly proportional to subjective monetary value.

Under the assumption of independence of the Gaussian distributions of the Thurstonian model (known as Case III), the equation for choosing option x over option y is:

$$P_{x \succ y} = \Phi\Big(\frac{V_T(x) - V_T(y)}{\sqrt{\sigma_x^2 + \sigma_y^2}}\Big) \tag{1}$$

It's easy to see that under the assumption of equal variances (Case V), this equation can become:

$$P_{x \succ y} = \Phi\Big(c * (V_t(x) - V_t(y))\Big) \tag{2}$$

Traditionally the Thurstonian model has been used to model items that are said to have a fixed subjective Thurstonian value Gaussian distribution, but in the case of this study the mean of this distribution can be free to shift under different contexts in the presence of context effects.

2. Experimental Procedure

30 participants were asked to respond to a series of likert and paired comparison questions given in random order. Items were generated in several different ways. 62 paired comparisons in both the probabilistic and temporal domains were chosen at random, with a smaller sooner and a larger later option in the temporal domain and a smaller but more likely versus a larger less likely option in the probabilistic domain. In order to gather information efficiently, lower delays were more likely to be chosen as options, and pairs were generated in a way that favored pairs likely to be judged to be similar in subjective value. 45 Pairs were chosen at random comparing probabilistic options with temporal options. In addition, 5 pairs in each of the three categories were chosen by the experimenter and were the same for each subject. Possible monetary rewards ranged from \$1 to \$4000, possible chances of reward ranged from 0.2 to 1, and possible delays ranged from 0 to 48 months.

In addition, participants were asked to compare 4 different monetary amounts at 4 different delays and 4 different probabilities with immediate certain rewards 3 different times. The immediate certain rewards were chosen randomly, but were bounded to be above amounts that the participant had chosen the probabilistic/delayed option over for the same discounted option or worse (lower or the same amount, less or the same delay/higher probability), and below amounts the participant had chosen when comparing to the same discounted option or better. Thus, as the experiment went on the possible range of immediate certain options shrunk to be more representative of the individual's true subjective value. Discounted amounts tested were at \$500, \$1000, \$2500, and \$4000, delayed at 3, 8, 18 or 48 months, or in the probabilistic case with chances of reward at 0.2, 0.4, 0.6, or 0.8.

Likert questions asked participants to rate on a scale from 0 to 9 (the keys on the keyboard) how good a particular delayed or probabilistic reward sounded. 55 delayed and 55 probabilistic items were chosen completely at random, with some chance for a certain or immediate reward, while 5 in each category were chosen by the experimenter and were the same for each individual. The response time for these likert questions was recorded. After each likert question, an additional likert question asked them to rate how difficult it was for them to assess the value of the item in question from 0 to 9. The point of these questions was to assess whether there might be evidence for differing degrees of uncertainty in subjective value across the attribute spaces to hint at whether a case III Thurstonian model should be used instead of a case V. After trying to fit a multivariate polynomial to individual percentile rankings of both reaction time and reported subjective difficulty of assessment, the R^2 was lower than 0.15 in both domains even with a 4th degree multivariate polynomial, so no evidence was found that there was a non-uniform amount of uncertainty across

both attribute spaces.

3. Modeling

The hyperbolic function was used to quantify people's subjective monetary equivalent for both delayed options and probabilistic options.

$$V_\$ = \frac{A}{1 + k_i(A) * D} \tag{3}$$

Describes the subjective monetary value for person i with delay discounting rate k_i at amount A. In other words, when a delayed option has a subjective monetary value of $V_\$$, an immediate monetary option greater than $V_\$$ is more likely to be preferred, while an immediate option less than $V_\$$ is more likely to be rejected in favor of the delayed reward. Higher values of k_i indicate that subject i deems monetary rewards to decrease in value more rapidly with respect to the magnitude of the delay.

Similarly:

$$V_\$ = \frac{A}{1 + h_i(A) * \theta} \tag{4}$$

Describes the subjective monetary value for person i with probabilistic discounting rate h_i at amount A, where $\theta = \frac{1-p}{p}$ is the odds against success.

Since both k and h have previously been found to vary with respect to monetary amount, they were modeled to vary linearly with respect to A:

$$k_i(A) = \beta_{k,i,0} + \beta_{k,i,1} * A \tag{5}$$
$$h_i(A) = \beta_{h,i,0} + \beta_{h,i,1} * A \tag{6}$$

Under the CAAH, the way two delayed options get discounted is slightly different than an immediate option with a delay. Specifically, drawing from the proposed functional form from Green et al. (2005) the sooner option gets discounted according to:

$$V_{\$,s} = \frac{A_s}{1 + w_{k,i} k_i(A_s) * D_s} \tag{7}$$

while the later option gets discounted according to:

$$V_{\$,l} = \frac{A_l}{1 + w_{k,i} k_i(A_l) * D_s + k_i(A_l) * (D_l - D_s)} \tag{8}$$

This can easily be extended to probabilistic discounting with the more probable option being discounted according to:

$$V_{\$,m} = \frac{A_m}{1 + w_{h,i} h_i(A_m) * \theta_m} \tag{9}$$

while the less probable option's subjective monetary value is:

$$V_{\$,l} = \frac{A_l}{1 + w_{h,i} h_i(A_l) * \theta_s + h_i(A_l) * (\theta_l - \theta_m)} \tag{10}$$

It should be noted that when w=1, or when D or θ are zero for one of the options, the CAAH reduces to the PVCH.

3.1. Linking the subjective monetary equivalent to the Thurstonian value

In the past, some other researchers have utilized the logistic function to study discounting (e.g. Luckman et al. (2017)). However, those that have done this did so with no transformation of the subjective monetary equivalent. This is problematic, since the comparison between options with subjective values equal to 1 USD vs 2 USD is likely to be much more substantial than 3000 USD vs 3001 USD, yet with no transformation this approach assumes the same probability of choosing the smaller vs the larger option. Also, the logistic function assumes a range of values from $-\infty$ to ∞, whereas here we are dealing with values from 0 to ∞.

It can be shown that the logistic function used here is the equivalent of what Luce's Choice Rule (Luce, 1959) would be under logarithmic transformation of the subjective monetary values. In the case of Luce's Choice Rule, the probability of selecting option x over option y is given by:

$$P_{x \succ y} = \frac{V(x)}{V(x) + V(y)} \tag{11}$$

It is a matter of personal opinion that the Thurstonian model, which assumes the evaluation of each option is Gaussian Distributed and the option deemed to be of higher value at that moment is a more plausible model. Either way, under a log transformation and utilization of the logistic function and an assumption of equal variance under the Thurstonian Model (Case V), the characteristics of the two sigmoidal functions are fairly similar and unlikely to yield dramatically different results.

The log transformation of $V_\$$ as being reflective of subjective value has a centuries old history going back to Daniel Bernoulli (Bernoulli, 1738). It has also been well known for over a century that many psychophysical stimuli have a subjective intensity that is proportional to the logarithm of the physical intensity according to the Fechner-Weber law (Fechner, 1860). There is also an advantage that taking the logarithm delivers the same results regardless of currency exchange rates. For this study, we assume that the Thurstonian Value is thus equal to the log transformation of the monetary subjective value:

$$V_T = log(V_\$) \tag{12}$$

Under the Case V Thurstonian Model this yields:

$$P_{x \succ y} = \Phi\Big(c_i(V_T(x) - V_T(y|x))\Big) \tag{13}$$

3.2. Bayesian Hierarchical Model

Although each individual's parameters were estimated individually, it was assumed that certain individual parameters were drawn from the same distribution.

3.2.1. Discounting Rates

Discounting rates and their rate of change were estimated hierarchically, with individual parameters for subject i being drawn as follows:

$$\beta_{k,i,0} \sim N(\mu_{k,0}, \sigma^2_{k,0})T(max(0, -\beta_{k,i,1} * 4000), \infty) \tag{14}$$
$$\beta_{k,i,1} \sim N(\mu_{k,1}, \sigma^2_{k,1}) \tag{15}$$
$$\beta_{h,i,0} \sim N(\mu_{h,0}, \sigma^2_{h,0})T(max(0, -\beta_{h,i,1} * 4000), \infty) \tag{16}$$
$$\beta_{h,i,1} \sim N(\mu_{h,1}, \sigma^2_{h,1}) \tag{17}$$

The truncation on the coefficient terms was placed to bound the discounting rate to always be positive.

The hyperparameters for the discounting rates were given semi-informative dis-

tributions as follows:

$$\mu_{k,0} \sim N(0, 1) \tag{18}$$
$$\mu_{k,1} \sim N(0, \frac{1}{1000}) \tag{19}$$
$$\mu_{h,0} \sim N(0, 10) \tag{20}$$
$$\mu_{h,1} \sim N(0, \frac{1}{400}) \tag{21}$$
$$\sigma^2_{k,0} \sim invgamma(0.5, 0.0002) \tag{22}$$
$$\sigma^2_{k,1} \sim invgamma(0.3, 3 * 10^{-11}) \tag{23}$$
$$\sigma^2_{h,0} \sim invgamma(0.4, 0.2) \tag{24}$$
$$\sigma^2_{h,1} \sim invgamma(0.6, 4 * 10^{-8}) \tag{25}$$

3.2.2. Testing the Common Aspect Attenuation Hypothesis

To test the CAAH for the probabilistic and temporal domains, the probability of person i having a $w_{k,i}$ or $w_{h,i}$ equal to one (indicating adherence to PVCH) was assumed to be equal for all individuals, but not necessarily equal for the probabilistic and temporal domains. These probabilities were both estimated with a uniform prior from 0 to 1. In the case that an individual was estimated to be adhering to the CAAH, $w_{k,i}$ and $w_{h,i}$ were both drawn from uniform distributions from 0 to 1. The Bayes Factor in the results section for each participant's adherence to CAAH was calculated according to the proportion of times $w_{k,i}$ or $w_{h,i}$ were not equal to one compared to the times they were.

3.2.3. Individual Sensitivity to Differences in Value

$c_{m,i}$ was given the following hierarchical prior:

$$c_{m,i} \sim gamma(a_m, b_m) \tag{26}$$
$$a_m \sim N(0, 10^9)T(0, \infty) \tag{27}$$
$$b_m \sim N(0, 10^9)T(0, \infty) \tag{28}$$

With m corresponding to either the probabilistic, temporal, or probabilistic with temporal domains.

3.2.4. Probabilistic Option vs Temporal Option

The analyses that did not have anything to do with the cross-domain questions were run without taking into account the cross-domain data. After it was concluded that the CAAH was being followed (see results), the cross-domain data was analyzed

under that assumption pertaining to the within-domain data.

The Cross-Domain model used assumed that people were comparing the present value comparisons of the temporal and probabilistic options with each other, but with an additive bias for the probabilistic option under the Thurstonian model that was estimated hierarchically for each individual. The equation for this is given below:

$$P_{pr \succ del} = \Phi\Big(c_{c,i}(V_T(pr) + B_{p,i} - V_T(del))\Big) \tag{29}$$

$B_{p,i}$ for subject i was estimated with the following hierarchical prior:

$$B_{p,i} \sim N(\mu_B, \sigma_B^2) \tag{30}$$
$$\mu_B \sim \ N(0, 10) \tag{31}$$
$$\sigma_B^2 \sim invgamma(0.01, 0.01) \tag{32}$$

3.2.5. Allowing For Error Trials

Since subjects were asked to answer a lot of questions back to back, it was assumed that they may have selected some on accident. For each trial, the model specified that person i answered randomly with probability r_i, and otherwise answered according to the Thurstonian model. Each r_i was given a Beta(2,100) prior.

4. Results

4.1. Common Aspect Attenuation Hypothesis

For each individual, a Bayes factor was calculated to determine the relative posterior probability of the CAAH ($w < 1$) vs the PVCH ($w = 1$) in both the probabilistic and temporal domains. Figure 1 below shows a histogram of the Bayes factors in the temporal domain for each individual. It was found that everyone had between moderate to strong support in favor of the CAAH, indicating a violation of stochastic transitivity under the assumptions of the model.

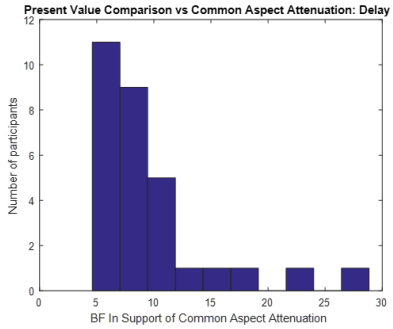

Figure 1:

In the probabilistic domain, there was a larger degree of variance in the individual Bayes factors, so the log Bayes factor in support of the CAAH is given in Figure 2. Although a couple people did not show evidence of response patterns favoring the CAAH over the PVCH, the overwhelming majority of participants showed strong to overwhelming support for CAAH, indicating a similar violation of stochastic transitivity under the assumptions of the model as was found in the delay case.

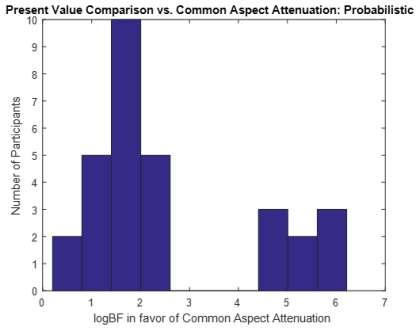

Figure 2:

4.2. Change in discounting rates with respect to monetary amount

For the most part, the results were consistent with previous research in finding that the majority of people decreased their discounting rate for larger monetary amounts for delayed rewards, and increased their discounting rate for larger monetary amounts for probabilistic rewards. This wasn't true for everyone, however. In fact, a few people in both domains had moderate evidence for having a change in discounting rate in the opposite direction, while a few did not show much change. A majority of people in both cases exhibited strong to overwhelming evidence in support of being consistent with the previous literature. The log Bayes Factors for individuals in both domains in favor of previous findings are presented in the two histograms below.

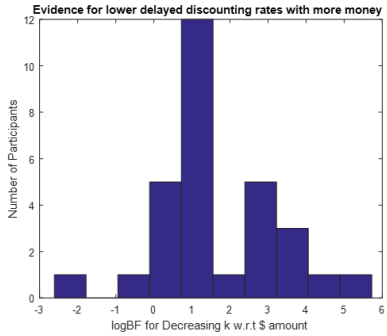

Figure 3:

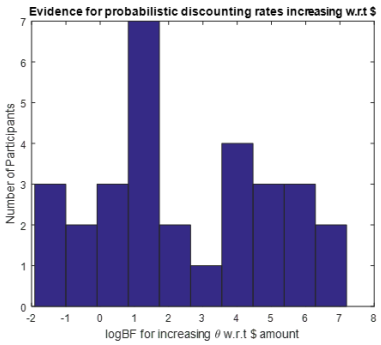

Figure 4:

4.3. Are people biased towards Delayed or Probabilistic options?

Looking at the posterior estimates for the bias term in favor of probabilistic options, we found that some people showed favor more towards the probabilistic option, while others showed more favor towards the delayed option, and some did not show any bias. The histogram of the individual posterior means for the additive pro-probability bias term is shown below.

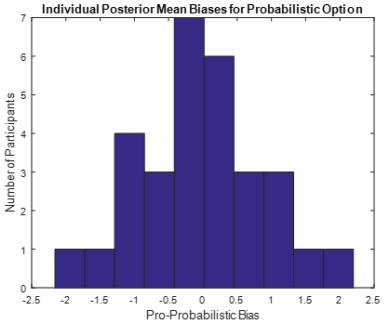

Figure 5:

Although for the group there was no consistent favoring of one side or the other, we find strong support for bias towards both probabilistic options and delayed options in certain individuals. The figure below shows the histogram of the posterior probability of having a positive pro-probabilistic bias. For some people, all of their posterior samples for the bias term were positive, while for others all of the samples were negative, and there were also many in-between. Given the assumptions of the model, this points to a context effect when comparing delayed with probabilistic options versus delayed with immediate or probabilistic with certain options in some individuals.

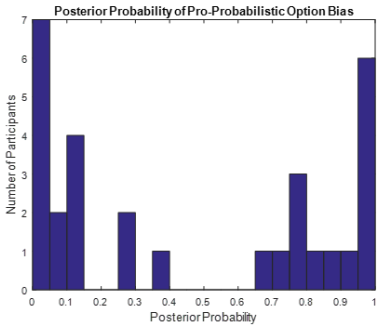

Figure 6:

5. Conclusion and Future Directions

The results of this experiment have largely confirmed findings from previous research regarding probabilistic and temporal discounting, as well as produced some new results. The results of this study point toward a number of different violations of stochastic transitivity in these domains at the individual level. At the same time, numerous assumptions have been made here regarding the functional form of the discounting functions, choice probabilities, and their behavior under different different contexts. A potentially fruitful future direction to take this research is to test these violations of individual stochastic transitivity in a manner that does not make assumptions on the functional form. It is possible to use a posterior predictive distribution from this experiment to find items that are most likely to elucidate these stochastic transitivity violations under a new experimental approach. This can not

only provide further evidence for these context effects, but also potentially point us in a direction that allows us to improve our understanding of the mathematical structure behind discounting.

6. References

References

Baker, F., Johnson, M. W., Bickel, W. K., 2003. Delay discounting in current and never-before cigarette smokers: similarities and differences across commodity, sign, and magnitude. Journal of abnormal psychology 112 (3), 382.

Bernoulli, D., 1738. Exposition of a new theory on the measurement of risk. Econometrica 22 (1), 23–36.

Du, W., Green, L., Myerson, J., 2002. Cross-cultural comparisons of discounting delayed and probabilistic rewards. The Psychological Record 52 (4), 479–492.

Fechner, G. T., 1860. Elemente der Psychophysik: Zweiter Theil. Breitkopf und Härtel.

Green, L., Myerson, J., 2004. A discounting framework for choice with delayed and probabilistic rewards. Psychological bulletin 130 (5), 769.

Green, L., Myerson, J., Macaux, E. W., 2005. Temporal discounting when the choice is between two delayed rewards. Journal of Experimental Psychology: Learning, Memory, and Cognition 31 (5), 1121.

Green, L., Myerson, J., McFadden, E., 1997. Rate of temporal discounting decreases with amount of reward. Memory & cognition 25 (5), 715–723.

Green, L., Myerson, J., Ostaszewski, P., 1999. Amount of reward has opposite effects on the discounting of delayed and probabilistic outcomes. Journal of Experimental Psychology: Learning, Memory, and Cognition 25 (2), 418.

Kirby, K. N., Petry, N. M., Bickel, W. K., 1999. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. Journal of Experimental psychology: general 128 (1), 78.

Luce, R. D., 1959. Individual choice behavior.

Luckman, A., Donkin, C., Newell, B. R., 2017. Can a single model account for both risky choices and inter-temporal choices? testing the assumptions underlying models of risky inter-temporal choice. Psychonomic bulletin & review, 1–8.

Myerson, J., Green, L., 1995. Discounting of delayed rewards: Models of individual choice. Journal of the experimental analysis of behavior 64 (3), 263–276.

Myerson, J., Green, L., Hanson, J. S., Holt, D. D., Estle, S. J., 2003. Discounting delayed and probabilistic rewards: Processes and traits. Journal of Economic Psychology 24 (5), 619–635.

Petry, N. M., 2001. Delay discounting of money and alcohol in actively using alcoholics, currently abstinent alcoholics, and controls. Psychopharmacology 154 (3), 243–250.

Rachlin, H., Raineri, A., Cross, D., 1991. Subjective probability and delay. Journal of the experimental analysis of behavior 55 (2), 233–244.

Thurstone, L. L., 1927. A law of comparative judgment. Psychological review 34 (4), 273.