Hierarchical Generalization Models for Cognitive Decision-making Processes

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Yun Tang, M.A., M.A.S.

Graduate Program in Psychology

The Ohio State University

2013

Dissertation Committee:

Professor Jay Myung, Advisor

Professor Mark Pitt

Professor Paulus De Boeck
Abstract

An important but challenging perspective of cognitive modeling is the generalization of a model. That is, whether a cognitive model can account for data observed under alternative experimental settings, either with new participants, under new experimental designs, or in different experimental tasks tapping the same cognitive processes. Most extant generalizability measures for model evaluation focus on the generalization of a model to a new participant sample, but within a given experimental setting. The present dissertation project develops a model evaluation method, dubbed as the hierarchical generalization models, to expand the current generalizability measures and assess the between-experiment generalization. This method utilizes the hierarchical Bayesian modeling approach to account for multilevel data combined from different experimental settings, to separate various sources of variability, and to identify generalizable model assumptions that are related to underlying cognitive processes.

This dissertation examines the applicability and plausibility of the hierarchical generalization modeling framework in the context of studies of behavioral decision-making. Two major experimental paradigms, the decision-from-description and decision-from-experience experiments, are extensively discussed with regard to the modeling of experimental data and the theoretical implication on the generalization of decision-making processes. The hierarchical generalization modeling framework demonstrates its
suitability for these decision-making paradigms through simulation studies and secondary data analyses.

Simulation studies in Study 1 and Study 2 demonstrate how to develop hierarchical generalization models to simultaneously model the experimental data from multiple decision-making paradigms. It also shows that hierarchical generalization models can appropriately recover the model structures of Cumulative Prospect Theory based models, which have been one of the mainstream theories in decision-making. Secondary data analyses in Study 3 and Study 4 further extend the application of hierarchical generalization models to heuristic-based stochastic models, and demonstrate the application in an empirical data set. The use of the hierarchical generalization models facilitates a comprehensive examination of the data, and provides evidence from a new, more integrated perspective, which adds to the support of ongoing discussions regarding the cognitive processes underlying the decision-making problem. These results, taken together, suggest that the hierarchical generalization modeling is a theoretically grounded and easily implementable approach to evaluate the ability of cognitive models to generalize across experimental tasks and designs.
To Professor Houcan Zhang and Professor Jay Myung
Acknowledgements

I would like to thank my advisor, Dr. Jay Myung, for the invaluable mentorship he provided all through my graduate study, and for being supportive and patient. I would like to thank my secondary advisor, Dr. Mark Pitt, for making this dissertation possible, and for his constructive advice and constant encouragement. I also wish to thank Dr. Paul De Boeck for his insightful comments and suggestions on the dissertation, and for the intellectually stimulating discussions through this year. Their guidance and engagement has made this dissertation a thoughtful and rewarding journey.

My sincere gratitude also goes to Dr. Woojae Kim, who was incredibly helpful in shaping the whole idea, and Dr. Dan Cavagnaro, who taught me so much about decision-making models. In addition, I am particularly indebted to the organizers of the Technion Prediction Tournament, for their generosity in publicly sharing the data.

The support and company of my friends has made the years in graduate school more enjoyable. I especially thank Miao Luo, Hyebin Rim, Nuo Xi, and Xin Yao for their six-year friendship. Finally, I am extremely grateful to my parents for their continuous love and support.
Vita

2004........................................B.S. Psychology, Beijing Normal University, China

2007........................................M.A. Psychology, Beijing Normal University, China

2012........................................Master in Applied Statistics, Department of Statistics,

The Ohio State University

2007 to present ............................Graduate Research and Teaching Associate,

Department of Psychology, The Ohio State University

Publications


Fields of Study

Major Field: Psychology

Area of Concentration: Quantitative Psychology
Table of Contents

Abstract .......................................................................................................................... ii
Acknowledgements ........................................................................................................ v
Vita ................................................................................................................................ vi
List of Tables .................................................................................................................. xi
List of Figures ................................................................................................................ xi
Chapter 1: Introduction ................................................................................................. 1
  1.1 From generalizability to generalization criterion ................................................. 1
  1.2 Hierarchical Generalization Models .................................................................... 9
    1.2.1 Basic ideas ................................................................................................... 9
    1.2.2 Operationalization of generalization assumptions ....................................... 12
  1.3 Purpose of study ................................................................................................. 16
Chapter 2: Decision-from-experience Experiments and Modeling ......................... 18
  2.1 Decision-from-experience paradigms .................................................................. 18
  2.2 Modeling decision-from-experience paradigms ............................................... 21
    2.2.1 Theoretical issues ....................................................................................... 21
    2.2.2 Models for decision-from-experience data ............................................... 24
    2.2.3 Model comparison and integration across tasks ......................................... 26
  2.3 Individual differences in decision-from-experience data .................................. 29
Chapter 3: Hierarchical Bayesian Modeling for Risky Choice Experiments ........... 32
  3.1 Hierarchical Bayesian modeling ......................................................................... 32
3.1.1 Graphical representation of hierarchical Bayesian modeling .......................... 35
3.1.2 Estimation of hierarchical Bayesian models ...................................................... 38
3.1.3 Model comparison ......................................................................................... 41
3.2 Hierarchical Bayesian modeling adaption of CPT .................................................... 43
3.3 Implementation of hierarchical generalization models ............................................ 48

Chapter 4: Part I – Simulation Study ............................................................................. 53
  4.1 Study 1: Parameter Recovery Study ...................................................................... 53
    4.1.1 Fully generalizable model ............................................................................. 54
    4.1.2 Partially generalizable model ...................................................................... 56
  4.2 Study 2: Model Recovery Study ........................................................................... 58
    4.2.1 Fully generalizable models .......................................................................... 61
    4.2.2 Partially generalizable models ...................................................................... 64
    4.2.3 Estimation of DIC differences ....................................................................... 67

Chapter 5: Part II – Secondary Data Analysis ................................................................. 72
  5.1 Technion Prediction Tournament .......................................................................... 72
  5.2 Models from TPT ................................................................................................. 74
    5.2.1 CPT-based models ....................................................................................... 76
    5.2.2 Natural-mean heuristic ................................................................................ 77
  5.3 Study 3: HGM comparison of CPT-based models ............................................... 81
    5.3.1 Data summary ............................................................................................... 81
    5.3.2 Models .......................................................................................................... 82
    5.3.3 Results and discussion .................................................................................. 84
  5.4 Study 4: HGM comparison between CPT and heuristics ...................................... 88
    5.4.1 Models .......................................................................................................... 88
List of Tables

Table 1. Examples of hierarchical generalization models assuming a constant model for Task 1 ................................................................. 15

Table 2. Pseudo code for obtaining the posterior samples for $\theta$ using conditional simulation .................................................................................. 40

Table 3. Pseudo code for obtaining the posterior samples for $\theta$ using Gibbs sampler .......... 40

Table 4. Summary statistics for means of hyper-parameter distributions in the fully generalizable model ........................................................... 56

Table 5. Summary statistics for means of hyper-parameter distributions in the partially generalizable model ......................................................... 57

Table 6. Competing fully generalizable models .............................................................. 62

Table 7. Estimates for DIC related values, and the difference in DIC values .................. 64

Table 8. DIC values estimated from fitting simulated data to the competing fully generalizable models ............................................................................. 64

Table 9. Competing partially generalizable models ........................................................ 65

Table 10. Estimates for DIC related values with data generated from Model TK-TK ..... 66

Table 11. Estimates for DIC related values with data generated from Model TK-Pow ... 66

Table 12. DIC values estimated from fitting simulated data to the competing fully and partially generalizable models ......................................................... 67

Table 13. Summary statistics for DIC differences of the problematic model pairs .......... 70
Table 14. Gamble designs and the proportion of choosing the risky prospect in DFD and DFE-S paradigms. ................................................................. 82

Table 15. Competing hierarchical generalization models for the selected TPT data sets. 84

Table 16. Estimates of DIC related values for the competing model set with variant parameter values (PV). ................................................................. 85

Table 17. Estimates of DIC related values for the competing model set with partially invariant parameter values (PPI). ................................................................. 85

Table 18. Competing hierarchical generalization models for the selected TPT data sets. 89

Table 19. Estimates of the DIC related values for the competing model set. ................. 91

Table 20. Pseudo code for bootstrapping process to estimate the DIC values. ............ 92

Table 21. Summary statistics for the bootstrap DIC differences between the models of CPT_Luce – NM_tr and CPT_tr – NM_tr. ................................................................. 92

Table 22. DIC values estimated from fitting simulated data to the models with the TK probability weighting function (N = 100, \( \sigma \mu \alpha = \sigma \mu \gamma = .01 \)). .......................... 117

Table 23. DIC values estimated from fitting simulated data to the models with the TK probability weighting function (N = 50, \( \sigma \mu \alpha = \sigma \mu \gamma = .01 \)). .......................... 117

Table 24. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function (N = 100, \( \sigma \mu \alpha = \sigma \mu \gamma = .01 \)). ......................... 117

Table 25. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function (N = 50, \( \sigma \mu \alpha = \sigma \mu \gamma = .01 \)). ......................... 117
Table 26. DIC values estimated from fitting simulated data to the models with the power probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{\gamma}} = .01$). .................. 118

Table 27. DIC values estimated from fitting simulated data to the models with the power probability weighting function (N = 50, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{\gamma}} = .01$). .................. 118

Table 28. DIC values estimated from fitting simulated data to the models with the linear probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{w}} = .01$). .................. 118

Table 29. DIC values estimated from fitting simulated data to the models with the linear probability weighting function (N = 50, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{w}} = .01$). .................. 118

Table 30. DIC values estimated from fitting simulated data to the models with the TK probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{\gamma}} = .01$). .................. 119

Table 31. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{\gamma}} = .01$). .................. 119

Table 32. DIC values estimated from fitting simulated data to the models with the power probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{\gamma}} = .01$). .................. 119

Table 33. DIC values estimated from fitting simulated data to the models with the linear probability weighting function (N = 100, $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{w}} = .01$). .................. 119
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Conceptual graphical model of a hierarchical model for multi-task data.</td>
<td>12</td>
</tr>
<tr>
<td>Figure 2</td>
<td>A simple hierarchical model in terms of a graphical model.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Graphical representation of a hierarchical Bayesian model implementing CPT.</td>
<td>46</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Representative shapes of four probability weighting functions.</td>
<td>50</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Graphical scheme of a partially generalizable p-FIPI model.</td>
<td>52</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Graphical scheme of a fully generalizable FIPI model.</td>
<td>52</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Distributions of hyper-parameter means estimated from the posterior sample in fitting the fully generalizable model.</td>
<td>56</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Distributions of hyper-parameter means estimated from the posterior sample in fitting the partially generalizable model.</td>
<td>58</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Histograms of the DIC difference for the three problematic model pairs.</td>
<td>71</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Graphical model for implementing the two-stage CPT model in the hierarchical Bayesian modeling framework.</td>
<td>77</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Graphical models for implementing the natural-mean heuristic in hierarchical Bayesian modeling.</td>
<td>79</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Posterior samples of hyper-parameters of the model TK-Linear with partially invariant parameter values (PPI), which has the smallest DIC value in Study 3.</td>
<td>87</td>
</tr>
</tbody>
</table>
Figure 13. Histogram of the bootstrap DIC differences between the CPT_Luce – NM_tr and CPT_tr – NM_tr models

Figure 14. Posterior samples of hyper-parameters of the model CPT_tr – CPT_tr

Figure 15. Posterior samples of hyper-parameters of the model CPT_tr – NM_tr
Chapter 1: Introduction

The ultimate goal of cognitive science is to understand the mental processes through which the human brain accomplishes complex tasks in everyday life, such as learning, memory, and problem solving. To this end, formal cognitive modeling provides useful language and tools to represent the underlying computational principles of human cognition. Cognitive psychology supplements experimentation with modeling and computer simulation (Friedenberg & Silverman, 2006). A cognitive model is a specific information-processing model of a mental ability, which quantifies the input and output of the process, describes the mechanism of the mental process, and formulates predictions about behavior. In other words, the goal of a cognitive model is to scientifically represent one or more of the basic cognitive processes, and depict how these processes interact (Busemeyer & Diederich, 2010). Cognitive models often are developed from a conceptual framework by translating its theoretical assumptions into a formal mathematical or algorithmic description. As cognitive modeling thrives, it is necessary to choose between different models (a.k.a. model selection) to assist model building and testing (e.g., Busemeyer & Diederich, 2010; Cavagnaro, Myung, & Pitt, in press; Myung & Pitt, 2001).

1.1 From generalizability to generalization criterion
Models may be evaluated and compared based on qualitative and quantitative criteria (Bussemeyer & Diederich, 2010; Myung, Tang, & Pitt, 2009). Qualitative criteria, such as interpretability, rely on knowledge, experience, and preferences of the modeler, and are subjective. On the other hand, quantitative criteria rely on the statistical aspect of modeling to evaluate models. These measures can quantify the value of a model, thus enabling explicit comparison among various models. The most often used quantitative criteria include the goodness-of-fit measures and the generalizability measures. Goodness of fit measures, such as $R^2$ and the root mean square error (RMSE), relate to how well the model fits the observed data, whereas generalizability measures relate to how well the model can predict future observations (Myung et al., 2009). Generalizability is currently the gold standard of model selection. Example measures of generalizability include: the Akaike Information Criterion (AIC); the Bayesian Information Criterion (BIC); the Normalized Maximum Likelihood (NML); and the Bayes Factor (BF), to name but a few.

Formally, generalizability is defined in terms of a discrepancy function that measures the degree of approximation or similarity between two probability distributions (Su, Myung, & Pitt, 2005). A discrepancy measure $D(f, g)$ between distributions $f$ and $g$ is an arbitrarily chosen function that satisfies $D(f, g) > D(f, f) = 0$ for $f \neq g$. Generalizability is then expressed as:

$$E^f_{\tau}[D(f_T, f_M)] \triangleq \int_y D\left(f_T(y), f_M(\hat{\theta}(y))\right) f_T(y)dy$$

(1)
where $f_T$ and $f_M$ denote the probability distributions of the true model and the model in consideration, $y$ denotes data, and $\hat{\theta}(y)$ is the best-fit parameter estimate. Equation 1 defines generalizability as a mean discrepancy between the true model and the best-fitting member of the model class of interest, averaged across all possible data $y$ that may be observed under the true model. According to this definition, the goal of model selection among a set of competing models is to select the one model that optimizes generalizability, or minimizes discrepancy.

Quantitative model evaluation criteria in general, and generalizability measures in particular, are useful in model development. Theories of generalizability measures and their applications in cognitive modeling abound in literature (e.g., Liu & Aitkin, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin, Lee, Kim, & Wagenmakers, 2008). One important constraint, however, is that the usefulness of these measures is limited to a specific experimental setting (Shiffrin et al., 2008). The experimental settings can refer to the experimental designs, and the experimental tasks conducted in a study. Generalizability is not directly observable in practice, and often is estimated from a data sample obtained in a specific cognitive task with a given experimental setting.

For the simplicity of explanation, what follows is explained in terms of experimental designs, though the same problem applies also to experimental tasks. Examples of experimental designs include the time interval between the stimulus and response in a psychophysical study, the word frequency in a retention experiment, and the length of reading materials in a learning task. The set of all values for $x$ that are allowed in an experiment is referred to as the design space of the model. The set of all realizations for $y$ that are possible under a certain model $f$ in an experiment is referred to
as the *data space* of the model. In practice, the realizations of $y$ are sampled from the data space under a certain experimental design $x$. Statistical models can be specified separately for $y$ conditional on $x$ in Equation 2, and for $x$ in Equation 3.

$$p_M(y, \theta|x) = p_M(y|x, \theta)p_M(\theta|x)$$ \hspace{1cm} (2)

$$p_M(x, \phi) = p_M(x|\phi)p_M(\phi)$$ \hspace{1cm} (3)

In the above specifications, the distribution of the experimental data $y$ given $x$ is parameterized with $\theta$, and the distribution of the design variable $x$ is parameterized with $\phi$. With most of the model selection methods, $x$ is assumed to be a known quantity and the focus is placed on the distribution of $y$ (Vehtari & Ojanen, 2012). In other words, model selection mostly concentrates on whether or not $p_M(y|x, \theta)$ is a satisfactory description of the data space.

The implicit assumption often is made that the experimental design $x$ is an appropriate representation of the design space. When this assumption is improbable, however, the term of $y$ in Equation 1 may be substituted with $y$ given $x$, where $x$ represents a sample from the design space of the experiment, i.e.,

$$\mathbb{E}^f[D(f, f_M)] \triangleq \int_{y|x} D \left( f(y|x), f_M \left( \hat{\theta}(y|x) \right) \right) f(y|x) d(y|x)$$ \hspace{1cm} (4)

Consider retention models in short-term memory, for example. A retention model describes a person’s ability to remember learned information after a short period of time.
In a retention experiment, the observed data $y$ is the proportion of study materials a person can recall, and the experimental design $x$ is the elapsed time between the study session and test session. Suppose that a retention experiment is conducted with a particular experimental design consisting of four retention intervals in seconds, denoted by $x = \{1, 5, 15, 30\}; y = \{y_1, y_5, y_{15}, y_{30}\}$ are the proportions of correctly recalled materials. A set of competing retention models are fit to $y$, and the generalizability measure is estimated for each model. A model $M_A$ with the smallest generalizability measure is then selected as the “best” model in the sense that newly observed $y$ from the same population is best predicted by $M_A$. Ideally, $M_A$ should be able to make accurate predictions for the proportions of recalled materials observed under other retention intervals, such as $x' = \{2, 12\}$. However, it is not theoretically guaranteed by most generalizability measures (e.g., AIC and BIC) because they consider only the known quantities of $x$ when estimating the generalizability measures. The constraint that generalizability measures are limited to a specific experimental design is undesirable for cognitive models, because cognitive models usually attempt to establish a generalized account for mental processes.

A solution for generalizability measures to reliably account for the design space is to formally include the design $x$ in Equation 1, which will give the following specification:

$$
E_f [D(f, f_M)] \triangleq \int_x \int_{y|x} D \left( f(y|x), f_M \left( \hat{\theta}(y|x) \right) \right) f(y|x)f(x) d(y|x) dx
$$

(5)
Note that the parameterization of the design space in Equation 3 is limited to certain types of variables, and may not be plausible for many cognitive experiments. Nonparametric solutions, such as the Indian Buffet Process (Griffiths & Ghahramani, 2011), are available to define probability measures over the design space, but they impose heavy implementational and computational costs.

An alternative way to approximate the design space is from a representative sample of designs. The generalization criterion proposed by Busemeyer and Wang (2000) is one attempt to treat the experimental design as a random sample from the design space, and evaluate the model’s ability to generalize to a different sample from the design space. Busemeyer and Wang (2000) examined the discrepancy between model predictions and a dataset collected under new designs (see also Forster, 2000), following an idea originated from the validity generalization criterion by Mosier (1951). The generalization criterion, as it is often called, involves a calibration design $x_1$, a generalization design $x_2$, and data observed under the different designs, $y$ given $x_1$ and $y$ given $x_2$. The measure for the generalization criterion can be expressed as follows:

$$E_f[D(f, f_M)] = \int_{y|x_2} D \left( f(y|x_2), f_M \left( \hat{\theta} (y|x_1) \right) \right) f(y|x_2) d(y|x_2)$$  \hspace{1cm} (6)

In the model selection procedure of the generalization criterion, the competing models are fit to data obtained from calibration design $y|x_1$, then the discrepancy function in Equation 6 is evaluated with data from generalization design $y|x_2$. The main advantage of the generalization criterion is that it takes a new sample from the design space, and
places emphasis on the new experimental design when estimating the generalizability measure. Hence, the generalization criterion attempts to evaluate a model’s ability to generalize by examining the model’s predictive performance on a new experimental design, which is the verbatim concern of generalization.

Busemeyer and Wang (2000) reported two examples to demonstrate the importance of considering the generalization criterion, in addition to generalizability measures, such as AIC. Example 1 is a hypothetical consumer preference study and the dependent variable consists of 500 levels of product utilities in an interval scale. The first 350 levels are used in the calibration stage and the last 150 levels are used in the generalization stage. Example 1 represents an extrapolation design because the generalization design is an extension of the calibration design. Example 2 is a secondary analysis of risky choice data collected in Busemeyer and Townsend (1993), with the original experimental design randomly partitioned into the calibration and generalization designs. In both examples, three competing models are considered, including a simple model, an overly complex model, and a principled model developed from reasonable theoretical assumptions. AIC-based model selection preferred overly complex models for both examples, whereas the generalization criterion preferred the principled models. In particular, Example 1 is a hypothetical example and the principled model is known to be closer to the data-generating model. The result that the generalization criterion preferred the principled model but not the overly complex model provided strong evidence that the generalization criterion has an advantage over some generalizability measures.

The generalization criterion has been applied in a few other model comparison studies. Ahn, Busemeyer, Wagenmakers, and Stout (2008) conducted sequential
decision-making experiments with two sets of clinically-validated designs (Iowa Gambling Task and Soochow Gambling Task), and compared the performance of two sequential decision models using the generalization criterion in short-term and long-term predictions. The result provided positive evidence for a model (Prospect Valence Learning model, PVL) from its predecessor (Expectancy Valence Learning model, EVL), which was consistent with other empirical studies supporting the PVL model (e.g., Fridberg et al., 2010; Yechiam & Busemeyer, 2008). Erev et al. (2010) also adopted the generalization criterion to compare decision-making models in terms of their predictive performance on a new set of experiment designs. A large set of models and various decision-making tasks were involved in Erev et al. (2010), and the results were not as clear-cut as other aforementioned studies. Yet the use of generalization criterion provides a new insight into the highly-debated modeling problem and additional evidence from the perspective of model selection.

However, the generalization criterion has limitations that hinder a wide usage. The need for partitioning designs implies more consideration in designing experiment and more effort and larger samples in collecting data. The procedure also requires a strong assumption on parameter invariance across the two designs, in that the predictions on the generalization design are obtained by the parameters fitted on the calibration design. If a theory has different explanations for different segments of the design space, and the calibration and generalization designs correspond to those different segments, the generalization criterion is likely to fail. For example, the cumulative prospect theory (CPT, Luce & Fishburn, 1991; Tversky & Kahneman, 1992) assumes different functional forms for gain and loss domains. If the criterion design includes only the gain domain and
the generalization design includes only the loss domain, the CPT model will not perform well in predicting for the generalization design and the subsequent model selection result will be misleading. Furthermore, as indicated in Equation 6, the generalization criterion does not consider the entire design space but is a mere extension of one sample of designs to another sample. The predictive performance on a third set of new designs still is not guaranteed. Therefore, the generalization criterion is a step forward to account for general experimental settings, but may not represent a full solution to the problem at hand, at least not yet.

1.2 Hierarchical Generalization Models

1.2.1 Basic ideas

The generalization criterion (Busemeyer & Wang, 2000) initiated a question that is important to cognitive modeling practice. Essentially, the generalization of a model is described as the model’s ability to predict what would be observed under altered circumstances (Shiffrin et al, 2008). It is desirable that cognitive models of interest generalize to new experimental designs, or to an alternative experimental task that shares the same mental processes. In other words, it would be useful to provide an evaluation method that extends the generalization criterion either to the entire design space or to the task space defined within a cognitive process, or to both. Hereafter, such a method is referred to as hierarchical generalization models. To simplify the discussion, this section will focus on generalizing a cognitive model to different experimental tasks while holding the experimental designs constant.
The general purpose of the hierarchical generalization models is to use a hierarchical modeling structure to check if a cognitive model can explain data observed in different experimental tasks in a parsimonious manner, or to help identify a generalizable structure in a cognitive model. The latter motivation would be particularly plausible for process models, of which the basic assumptions may be examined separately (e.g., Yechiam & Busemeyer, 2005). For the ease of explanation, the following notations from Draper (1995) will be adopted. The uncertainty about unknown quantities (i.e., dependent variables) $y$, based on known quantities (i.e., independent variables), $x$ is based on a model $M$ that formalizes assumptions about how $x$ and $y$ are related. $M$ consists of two components: structural assumptions $S$, such as functional forms and error distributions, and parameters $\theta$ associated with a given choice of $S$. The model selection problem is then to specify a plausible single "best" choice of $S^*$ for $S$, and proceed as if $S^*$ were known to be correct. For two tasks, assume $x_1, y_1$, and $M_1 = \{S_1, \theta_1\}$ for Task 1. $S_1$ may be further specified as a set $\{S_{11}, S_{12}\}$ associated with $\{\theta_{11}, \theta_{12}\}$. Terms for Task 2 are similarly notated as $x_2, y_2$, and $M_2 = \{S_{21}, \theta_{21}, S_{22}, \theta_{22}\}$.

Hierarchical generalization models can be implemented by combining $\{x_1, x_2, y_1, y_2\}$ to form multilevel data, model the data with a set of hierarchical models, and perform model selection to choose the best model. There are two key characteristics of the proposed approach. One is to combine experiment variables and observed data from multiple tasks. The other is the hierarchical model structure to incorporate task differences and individual differences.

Combining multiple experiment tasks can produce multilevel data, in which individual data $\{x_1, x_2, y_1, y_2\}$ consist of the lower level and are nested within the two
tasks at the higher level. Hierarchical modeling is a specifically appropriate approach for multilevel data. Some examples of hierarchical models include hierarchical linear models (e.g., MacCallum, Kim, Malaskey, & Kiecolt-Glaser, 1997) and hierarchical Bayesian models (e.g., Griffiths, Kemp, & Tenenbaum, 2008; Lee, 2008; Rouder & Lu, 2005; Shiffrin et al. 2008). In hierarchical models, parameters at the lower level are determined in part by other parameters at the higher level (a.k.a. hyper-parameters). A conceptual graphical model for the multiple-task data is illustrated in Figure 1. The parameters depicted in the model include task-specific parameters \{θ_{11}, θ_{12}, θ_{21}, θ_{22}\} and a hyper-parameter ϕ_1 which determines \{θ_{11}, θ_{21}\}. Structural assumptions S are not explicitly shown in Figure 1 but are implied by the parameters. That is, two of the task-specific structures \{S_{11}, S_{21}\} are determined by a hyper-structure S_1. As in the example, tasks are considered as the higher level in the hierarchical modeling of multiple-task data.

Different tasks can have the same structural assumption and shared hyper-parameters, with similar or different numerical values for task-specific parameters (e.g., θ_{11} and θ_{21}). Simultaneously, one or more tasks can have additional unique structures and parameters (e.g., θ_{12} to Task 1).
1.2.2 Operationalization of generalization assumptions

It can be useful to explicitly conceptualize the levels of generalization assumptions in the framework of hierarchical generalization models and match the assumptions with competing models. For all the assumptions, the concept of generalization is considered as mathematically equivalent to the concept of invariance. That is, if structure S or parameter \( \theta \), holds invariant in different tasks or different designs, it is equivalent to say that S or \( \theta \) is generalizable in the different experiment settings. Four levels of invariance are considered. Similar invariance assumptions are applied to individual differences in cognitive models (Kim, 1998).

The first and strictest form of generalization is “invariance in functional form and parameter value” (FIPI). It is assumed that every individual has the same structural assumption (\( S_1 = S_2 \)) and the same numerical value of parameters (\( \theta_1 = \theta_2 \)) across tasks, which produces the same predicted responses. In this case, the model should produce exactly the same response for a particular stimulus in each task. Any observed variation...
in data under the same experimental design is considered as random noise or sampling error. In other words, the realizations of $y_1$ and $y_2$ should be equivalent given $x_1 = x_2$. If a model for two tasks is invariant in value, it implies that the model is fully generalizable from one task to the other.

The second and less strict form of generalization is “invariance in functional form but variance in parameter value” (FIPV), which assumes that the functional forms in structural assumptions are invariant across tasks ($S_1 = S_2$), but the numerical values of parameters may vary ($\theta_1 \neq \theta_2$). Observed variation in data under the same experimental design could be due to random noise or to differences in parameter values. If a model is invariant in functional form, it can imply that same mental processes are involved in both tasks, although the processes act differently. It would be desirable to have theoretical justifications for the variation of parameters, but the model is nevertheless generalizable and the assumption of underlying processes is supported.

The third and even less strict form of generalization is “partial invariance in functional form and variance in parameter values” (p-FIPV). It may be explained with the help of the conceptual diagram in Figure 1. The partial invariance assumes that a subset of structural assumptions are invariant across tasks ($S_{11} = S_{21} = S_1$), while task-specific structural assumptions (e.g., $S_{12}$ and $S_{22}$) exist, pending theoretical justification. Consequently, each task would have its unique parameters ($\theta_{12}$ and $\theta_{22}$) as well as invariant parameters ($\theta_1 = \theta_2$) or hyper-parameter ($\phi_1$). In this case, there would be relatively large variation in data observed under the same experimental design for two tasks. Partial invariance in functional form implies a weak form of generalization. The predictive performance of the model from one task to the other would be very poor.
because of the unknown sources of variation associated with the new task. This level of invariance can nonetheless provide some evidence supporting a subset of generalizable mental processes.

The fourth and weakest form of generalization is “no invariance”, or “variance in functional form and parameter value” (FVPV), which assumes that all functional forms vary across tasks. The two tasks have completely independent structural assumptions $S_1$ and $S_2$, implying that neither $M_1$ nor $M_2$ generalizes to the new task. This class of invariance does not hold much theoretical interest in terms of generalization, except that it can serve as the null model when performing the generalization test.

The framework of hierarchical generalization models may be explained using a simple example of regression models. In this context, $x_1$ and $x_2$ are explanatory variables for each task, $y_{i1}$ and $y_{j2}$ are predictors, and $i$ and $j$ stand for individual data from each task. Assume that $M_1$ is a Bayesian linear regression model: $M_1: y_{i1} \sim \mathcal{N}(\beta_{i1} x_1, \sigma_{i1}^2)$. $S_{11}$ is a linear term with coefficient $\beta_{i1}$, and $S_{12}$ is normal error with standard deviation of $\sigma_{i1}$. $\beta_{i1}$ and $\sigma_{i1}$ have distributions $\beta_{i1} \sim \mathcal{N}(\mu_{\beta1}, \sigma_{\beta1})$ and $\sigma_{i1} \sim \mathcal{N}(\mu_{\sigma1}, \sigma_{\sigma1})$. Different model assumptions in $M_2$ will be associated with the four forms of generalization as summarized in Table 1. For example, if there is no invariance in functional forms or parameter values (i.e., FPI), $M_2$ should be very similar to $M_1$, i.e., $M_2: y_{j2} \sim \mathcal{N}(\beta_{j1} x_2, \sigma_{j1}^2)$, with $\beta_{j1} \sim \mathcal{N}(\mu_{\beta1}, \sigma_{\beta1})$ and $\sigma_{j1} \sim \mathcal{N}(\mu_{\sigma1}, \sigma_{\sigma1})$. If $M_2$ is a Bayesian log normal model: $M_2: y_2 \sim \ln \mathcal{N}(\gamma_{j2} \ln x_2, \tau_{j2}^2)$, where $S_{21}$ is a natural logarithmic term with coefficient $\gamma_{j2}$, and $S_{22}$ is log normal error with standard deviation of $\tau_{j2}$, with $\gamma_{j2} \sim \mathcal{N}(\mu_{\gamma2}, \sigma_{\gamma2})$ and
There is no generalizable assumption across Tasks 1 and 2 (i.e., FVPV).

Examples of FIPV and p-FIPV models are listed in Table 1.

Table 1. Examples of hierarchical generalization models assuming a constant model for Task 1.

<table>
<thead>
<tr>
<th>Models</th>
<th>Task 1</th>
<th>Task 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_{11}$</td>
<td>$S_{12}$</td>
</tr>
<tr>
<td>FIPV</td>
<td>$\beta_{11}x_1$ $N(0, \sigma_{11}^2)$</td>
<td>$\beta_{12}x_2$ $N(0, \sigma_{12}^2)$</td>
</tr>
<tr>
<td>FIPV</td>
<td>$\beta_{11}x_1$ $N(0, \sigma_{11}^2)$</td>
<td>$\beta_{12}x_2$ $N(0, \sigma_{12}^2)$</td>
</tr>
<tr>
<td>p-FIPV</td>
<td>$\beta_{11}x_1$ $N(0, \sigma_{11}^2)$</td>
<td>$\beta_{12}x_2$ $N(0, \sigma_{12}^2)$</td>
</tr>
</tbody>
</table>

The different invariance assumptions elaborated above also lead to the use of different hierarchical models, especially hierarchical Bayesian models. The evaluation of generalization in HGM is operationalized as a regular model selection procedure with the competing models carrying different levels of generalization assumptions. Using quantitative model selection methods, the extent to which a model generalized may be mapped onto an ordinal-like scale with the above four levels of invariance. If AIC is used and the best performed model shows partial invariance in functional form (p-FIPV), it can be interpreted that the model $M$ in consideration has a weak ability to generalize across tasks. This study concentrates on developing HGM as a model testing scheme. Model comparison results will be explained in a more descriptive way and no new quantitative measures will be developed.
1.3 Purpose of study

Generalization is a critical concept in theory development, and should be of major concern in cognitive modeling. It is an important property for the structural assumptions of a cognitive model to be consistent across various experimental settings, either with new experiment designs or in different experiment tasks bearing similar mental processes. The present dissertation is interested in the development of a new statistical method that allows to measure the ability of a cognitive model to generalize beyond the current experimental designs and task. The proposed method of hierarchical generalization models will extend the current measures of generalizability to include data observed from multiple tasks, as well as to test for generalization by examining the invariance of parameters and functional forms across tasks.

Hierarchical generalization models can provide a flexible framework accommodating nested or non-nested, Bayesian or frequentist-style models and can be combined with various model selection measures. In the current project, the hierarchical Bayesian modeling framework will be utilized to set up the hierarchical generalization models. The usefulness of the hierarchical generalization models will be demonstrated in the context of behavioral decision-making with a special focus on decision-from-experience experiments. In what follows, decision-from-experience experiments and the related theoretical questions are introduced in Chapter 2. Chapter 3 briefly summarizes the benefits of using the hierarchical Bayesian modeling framework, and continues to introduce how to set up the hierarchical Bayesian modeling for behavioral decision-making experiments and for the hierarchical generalization models.
After outlining the framework of hierarchical generalization models and the contextual background for the project, Chapter 4 describes a series of simulation studies to examine the plausibility and soundness of the proposed framework of hierarchical generalization models. Specifically, Study 1 checks if the hierarchical Bayesian modeling implementation of hierarchical generalization models can account for the decision-making data by correctly recovering the functional and parameter structure of the model in consideration. Study 2 focuses on whether the Deviance Information Criterion (DIC) can be appropriately integrated into the framework to facilitate the comparison of the competing hierarchical generalization models. Chapter 5 discusses another approach of secondary analysis to examine the framework of hierarchical generalization models. An empirical data set obtained from behavioral decision-making experiments is analyzed using the hierarchical generalization models. The implications of the hierarchical generalization models for building up or modifying cognitive models are explored. In particular, emphasis is placed on how the theoretical question for generalization can be realized in the hierarchical generalization models.
Chapter 2: Decision-from-experience Experiments and Modeling

The framework of hierarchical generalization models is demonstrated in the context of studies of behavioral decision-making. This chapter introduces two major experimental paradigms that are commonly employed in the decision-making studies, namely, the decision-from-description and decision-from-experience experiments. In particular, the paradigms in the decision-from-experience experiments, and the related theoretical questions are elaborated. Discussion of the theoretical questions motivates the application of hierarchical generalization models.

2.1 Decision-from-experience paradigms

Many decision-making studies, including the well-known prospect theory studies (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), make regular use of hypothetical monetary gambles as a medium and assume that many real-world events resemble the same properties of gambles, having several outcomes and associated probabilities (Hertwig & Erev, 2009). The typical implementation of the monetary gambles usually states the outcomes (gains or losses) and associated probabilities for two or more options and requires decision-makers to indicate their preference among the options. This kind of monetary gamble can be termed decision-from-description, as the decision-makers make choices based on the full information of the payoff distributions for available options. To give a concrete example, suppose that a monetary gamble offers
two options: gamble A has a probability of 90% to win $1 and 10% to win $10, and
gamble B has a probability of 70% to win $5 but 30% to lose $5. In such decision-from-
description experiments as this, decision-makers have full access to all information of
possible outcomes and associated probabilities.

Other researchers (e.g., Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev,
2004) argue that, in everyday life, there are situations in which people cannot obtain full
descriptions of probability and outcome information but have to rely on their past
experience or learn such information in a trial-and-error manner. For example, when
deciding how to commute to school after moving, one may need to try different routes on
different days before choosing one that is satisfactory. In contrast to description-based
decisions, the latter case is usually referred to as experience-based decision or decision-
from-experience. Decision-from-experience is of interest for various reasons. The
simplified decision-making scenario in experience-based experiments offers an even
closer resemblance to many daily life decision scenarios. It also offers an experimental
medium through which dynamic processes in decision-making, especially information
integration and learning, are studied.

Experimental studies using the decision-from-experience paradigm have
increased in the recent decade (for a review, see Hertwig & Erev, 2009). Decision-from-
experience experiments usually present the decision-maker with two or more options in
the form of card decks (or buttons etc.) shown on a computer monitor. Each deck
represents a hidden payoff scheme similar to the fully described ones used in the
decision-from-description experiments. The experimental tasks vary in the literature but
can be roughly categorized into two major paradigms. For illustration purpose, assume
that a gamble pair is used in the decision-from-experience experiment, of which deck A has a flat rate of always winning $2, whereas deck B has a probability of 90% to win $1 and 10% to win $10. This gamble pair is used in all the paradigms discussed below.

In the *sampling paradigm* (e.g., Hertwig et al., 2004; Weber, Shafir, & Blais, 2004), a decision-maker freely chooses between the two decks, observes the outcome she might earn, and chooses one deck when she is willing to make the final decision. The actual outcome of the gamble is determined by the payoff scheme of the chosen deck. Say a person has chosen A-A-B-B-A, observed $10-$1-$2-$2-$1, indicated deck A for her final choice, and had $1 for this final outcome; she would earn $1 from this game.

In the *repeated-choices paradigm* (e.g., Barron & Erev, 2003; Erev & Barron, 2005), a decision-maker makes sequential choices from either deck, observes and accumulate the outcome associated with the chosen deck, and continues after a certain number of choices determined by the researcher. A decision-maker has no access to non-chosen decks. If a game stopped at the fifth choice, the person has chosen A-A-B-B-A, and had $10-$1-$2-$2-$1 for outcomes; she would earn a total of $16 from this game. The repeated-choices paradigm can be extended to include full feedback (e.g., Jessup, Bishara, & Busemeyer, 2008; Yechiam & Busemeyer, 2006), in which a decision-maker observes the outcome for the non-chosen option. However, there are relatively few empirical studies using the full-feedback paradigm and they are not be discussed at length.

Decision-from-experience paradigms attract much research interest, mostly because systematic and obvious behavioral deviations from decision-from-description paradigm, and a long debate discussing the need for “different theories of risky choice”
was initiated (Hertwig et al., 2004; Fox & Hadar, 2006; Rakow & Newell, 2010; Erev et al., 2010). Another major motivation for the decision-from-experience studies is that the decision-from-experience paradigms necessarily involve many cognitive processes such as learning, memory, knowledge representation, acquisition and integration of information, and choice mechanism. Consequently, studying the decision-from-experience paradigms can facilitate the modeling of human decision-making processes, which further provide an integrated theory for decision-making phenomena (Weber & Johnson, 2009).

2.2 Modeling decision-from-experience paradigms

2.2.1 Theoretical issues

One of the most debated problems in decision-from-experience study is the gap between description and experience paradigms (the description-experience gap, e.g., Erev et al., 2010; Hertwig & Erev, 2009; Rakow & Newell, 2010; Weber et al., 2004). The most discussed form of the description-experience gap lies in weighting the probability of rare events. According to prospect theory, people overweight lower probability events and underweight higher probability events (Kahneman & Tversky, 1979). These patterns are repeatedly observed in description-based paradigms (for a review, see Weber et al., 2004). In the experience-based paradigms, however, this pattern is reversed, and rare events have generally less impact than in the description paradigm. The underweighting of rare events is considered one of the major differences underlying the description-experience gap (Erev & Barron, 2005). Therefore, this key phenomenon requires consideration in the successful modeling of the decision-from-experience paradigms.
Two behavioral patterns in decision-from-experience experiments that are closely related to the underweighting of rare events are sampling bias and recency effect. Sampling bias usually is exhibited in decision-from-experience experiments; for example, a subject tends to draw a relatively small number of samples to learn about the payoff distribution before making a final decision. As a result, the designated payoff distribution is incorrectly perceived, and the probability of each outcome is misunderstood. Some researchers believe that the description-experience gap is attributed to the sampling error during experiments. In earlier decision-from-experience studies (e.g., Hertwig et al., 2004; Weber et al., 2004; see also Hau, Pleskac, Kiefer, & Hertwig, 2008), participants drew a small sample, which systematically distorted the encountering of rare events. Fox and Hadar (2006) claimed that the participants’ choices were similar to the choices predicted from prospect theory when using experienced probability as the input. Consequently, they suggested that, other than investigating the probability weighting process, future decision-from-experience studies should instead explore models of search rules, search termination, and bias in likelihood judgment. On the other hand, many studies have attempted to use various approaches to increase a decision-maker's sampling effort, or to match the experienced probabilities during the experiment (Hau et al., 2008; Newell & Rakow, 2007; Rakow, Demes, & Newell, 2008). Results generally reveal that more search and better observation of the payoff distribution have reduced the description-experience gap, but cannot eliminate it. In short, sampling error is an important factor to consider when modeling decision-from-experience paradigms, but more processes can be involved and contribute to the description-experience gap, even after accounting for the bias in probability.
Recency effect is another behavioral pattern that is commonly associated with underweighting of rare events and is believed to contribute to the description-experience gap. Empirical results show that later observations of the choice options predict final choice better than earlier observations. Recency effect is usually explained as a memory component in decision-making models and heuristics for the decision-from-experience paradigms (for brief examples of its usage, see Hau et al., 2008; Gonzalez & Dutt, 2011). However, other studies show that the recency effect does not emerge consistently (e.g., Ungemach, Chater, & Stewart, 2009). According to Lejarraga and Gonzalez (2011), different models can have distinct assumptions about how memory affects decisions in the decision-from-experience paradigms. For example, instance-based learning models assume that all outcomes are stored in memory and more recent outcomes have a larger influence (i.e., recency effect is assumed). On the other hand, reinforcement learning models assume that expectancy is stored in memory after being updated from recent outcomes, but outcomes are not stored (i.e., no explicit recency effect is assumed). The different model assumptions can still result in similar model fit or choice predictions, therefore increasing the difficulty in theory testing.

Other factors that may affect the underweighting of rare events include the representation distortion in the decision-from-experience paradigms (Camilleri & Newell 2009), the complexity of problem scenario, and the individual difference in numeracy (Lejarraga, 2010). Again, these sources of effects have confounded the investigation of the decision-making mechanism underlying the decision-from-experience paradigms. This is the place in which quantitative modeling can provide a rigorous method to dissociate and test the effects from different processes.
2.2.2 Models for decision-from-experience data

For the various cognitive processes involved in the decision-from-experience paradigm, quantitative modeling seems a popular approach to account for the behavior patterns in these experiments. In recent years, much effort has been devoted to the modeling of the decision-from-experience experiments. There are roughly three categories of models that receive significant attention in modeling the decision-from-experience experiments, namely, extensions of CPT, heuristics, and learning models. Many of these models are examined in Hau et al. (2008) and Broomell, Budescu, and Por (2011), and exhibit generally good predictive performance.

Firstly, extending CPT models seems to be a natural choice for modeling the decision-from-experience experiments, especially given that many experience-based tasks use monetary gambles with two prospects, as in the classical description-based paradigm. Fox and Hadar (2006) proposed a two-stage CPT model to specifically address the issue of sampling bias. In the two-stage CPT model, the exact payoff distribution (i.e., experienced probability and observed outcomes) is first obtained, and then the value function and probability weighting function are applied to account for the decision maker's choice. The two-stage CPT model typically relies on the parameter estimates from Tversky and Kahneman (1992) or similar sources. In other words, the parameters for probability weighting function reflect a similar trend, as in the decision-from-description experiments. Namely, smaller probabilities are overweighed and larger probabilities are underweighted. There are also other variations of CPT models considered for decision-from-experience data. For instance, Ann and Picard have
implemented the CPT model with aspiration levels in the Technion Prediction Tournament (see Erev et al., 2010).

Secondly, the heuristic approach to decision-from-experience data is also an important and widely studied approach in the decision-from-description paradigm (Gigerenzer & Gaissmaier, 2011). In the modeling of the decision-from-experience data, many heuristics are also considered and tested (Hau et al., 2008). Among these heuristics, the maximax and natural mean heuristics outperform many other heuristics. In addition to the application of single heuristics, an ensemble of various heuristics also showed good predictive performance. For example, Herzog, Hau, and Hertwig implemented four rules together to predict the decision-from-experience sampling paradigm in the Technion Prediction Tournament (Erev et al., 2010), including variations from the natural mean heuristic (Hertwig & Pleskac, 2008), a stochastic version of CPT, and a stochastic version of priority heuristic (Rieskamp, 2008).

The third category of learning models is exclusive to decision-from-experience paradigms as it considers the dynamic nature that is a characteristic feature in the experience-based paradigms but not in the description-based paradigms. Reinforcement learning models are the most representative in this category. Reinforcement learning refers to the learning of how to map situations to actions to maximize reward (Sutton & Barto, 1998). It is an area of machine learning that focuses on goal-directed learning from the learner’s interaction with the environment. Reinforcement learning models are used more frequently in the decision-from-experience repeated-choice paradigm. Examples include the explorative sampler model and its variations (Erev, Ert, & Yechiam, 2008b; Erev & Haruvy, 2009), and extensions of the expected valence learning model (EVL,
Busemeyer & Stout, 2002; for adaptations to the decision-from-experience paradigms, see Koritzky & Yechiam, 2010; Yechiam & Busemeyer, 2006). Another flavor of learning model for the decision-from-experience data is based on the Instance-Based Learning Theory (IBLT, Gonzalez, Lerch, & Lebiere, 2003). IBLT is a learning theory proposed for dynamic decision-making. It suggests that a decision-maker learns with the accumulation and refinement of instances, such as context and choices. IBLT has been applied to the sampling and repeated-choice paradigms of decision-from-experience experiments (Gonzalez & Dutt, 2011; Lejarraga & Gonzalez, 2011), and was claimed to be a generalized theory to account for the decision-making mechanism in the decision-from-experience experiments.

One characteristic in developing various quantitative models for experience-based paradigms is the explicit incorporation of cognitive processes in model hypotheses, so that the models are connected to the theoretical implications revealed in decision-from-experience experiments. For instance, underweighting of rare events can be directly expressed in CPT by setting the parameters for probability weighting functions ($\gamma$ and $\sigma$) larger than 1. The resulted probability weighting curve then changes to an S shape instead of an inverse-S shape, suggesting that events with smaller probabilities (i.e., rare events) are underweighted and events with larger probabilities are overweighed. Alternatively, underweighting of rare events can be indirectly modeled through a memory-like component. Such a component samples a subset of experienced outcomes, estimates a potentially biased probability, and estimates the utility of prospects from the estimated probability.

2.2.3 Model comparison and integration across tasks
There are two important issues to consider in the modeling of the decision-from-experience experiments. Firstly, as different models for the decision-from-experience data attempt to explain the underlying decision-making mechanism through different model hypotheses or cognitive process, model comparisons are needed to determine which model can better describe empirical data and explain the underlying mechanism. However, the proposed models are not easily distinguishable. Models or model classes with diverse assumptions usually exhibit similar performance in model fit to experiment data or in a predictive validity check (Hau et al., 2008; Erev et al., 2010; Gonzalez & Dutt, 2011). Various model comparison criteria and procedures have been applied to the decision-from-experience data, including the mean square distance (MSD, Erev et al., 2010), the Akaike information criterion (AIC, Gonzalez & Dutt, 2011), the Multidimensional Scaling (MDS) of pair-wise comparisons (Broomell et al., 2011), and the generalization criterion (Erev et al., 2010). None of these comparisons, however, has yielded clear-cut results so far. Although similar groups of models have been compared, model preference by different model comparison tends to vary across studies or data sets (e.g., Gonzalez, Dutt, & Lejarraga, 2011). The search for an efficient and meaningful model comparison method would be beneficial for the modeling of experience-based decision.

Secondly, the other central question in modeling the decision-from-experience paradigms is whether different theories are required to explain description- versus experience-based decisions. There has been a continuing debate whether “different theories of risky choice” (thus, different models) are needed for each paradigm, including the description-experience gap and the variations within the decision-from-experience
paradigms (Hertwig et al., 2004; Fox & Hadar, 2006; Rakow & Newell, 2010; Erev et al., 2010). Data observed in different paradigms are better explained by different models (Erev et al., 2010). In comparing different models, it could be preferred to place more weight on the model’s ability to generalize. In other words, for a certain model, if parameter estimates from one decision-from-experience paradigms could well predict the behavioral pattern on another decision-from-experience paradigm, this model should be valued even if its model fit might be inferior to some other model. Therefore, the first step toward integrating the decision-from-experience studies should be to identify a model or some structural assumption that generalizes across different decision-from-experience paradigms. Recently, Gonzalez and Dutt (2011) demonstrated that a single cognitive model based on the instance-based learning theory (IBLT, Gonzalez et al., 2003) captures the choice and behavioral pattern in both sampling and repeated-choice paradigms of decision-from-experience experiments. They showed that the IBLT outperforms the best models created for each paradigm separately (from Erev et al., 2010) using the AIC as the criterion. Although the IBLT has its flaws in accounting for individual data (Hills & Hertwig, 2012), it demonstrates an important advance in integrating various decision-from-experience paradigms and starts working toward filling the description-experience gap.

To summarize, the diversity of the decision-from-experience paradigms in the literature has contributed to the current confusion in modeling decision-making behavior. Approaching the problem from a hierarchical generalization modeling perspective of generalization may provide new insight into the ongoing debate. In particular, hierarchical generalization modeling can shed light on identifying generalizable
assumptions within various decision-from-experience paradigms, and help building an integrated theory for empirical results.

### 2.3 Individual differences in decision-from-experience data

Another much less discussed issue in modeling the decision-from-experience experiments is the neglect of different sources of variability. To examine the differences in different paradigms (e.g., decision-from-description, decision-from-experience sampling, and decision-from-experience repeated-choice), various sources of variability are entangled, including but not limited to participant variability, gamble payoff design variability, and task variability. The importance of individual differences in the decision-from-experience experiments is worth special discussion.

It is desirable to fit models to individual data and consider individual parameters in making inferences (Ahn et al., 2008). A common limitation in many models used in the decision-from-experience paradigm (Erev et al., 2010; Gonzalez & Dutt, 2011; Hau et al., 2008) is the neglect of individual difference by pooling the data and performing a group-level analysis. There can be theoretical and technical reasons for the pooling. It may be a reasonable assumption that each individual utilizes similar decision-making mechanisms. Averaging the data could be advantageous in canceling out random error. Furthermore, due to the small number of data points, which is especially common in the repeated-choice paradigms, some models are not identifiable at the individual level (Hau et al., 2008).

Whereas pooling the experimental data has some advantages, the neglect of individual differences can sometimes hurt a model’s inference in explaining the
behavioral pattern of a decision-maker. A case in point is Gonzalez and Dutt (2011), who claimed that the alternating rate (A-rate) in sampling different prospects decreased as the number of samples increased. Hill and Hertwig (2012), however, pointed out that this trend was likely to be an artifact because individual A-rates had been aggregated inappropriately. Both the number of samples and the A-rate were positively skewed with a negative correlation between the two. Because the aggregation was not normalized according to the frequency of samples, those individuals who sampled fewer but alternated more were more likely to account for a smaller part in the final analysis, and those individuals who sampled more and alternated less appeared to be more salient to the right tail. Consequently, the artifact arises that exploration gradually changed to exploitation in the sampling paradigm. Hill and Hertwig (2012) calculated a normalized aggregated A-rate and showed that the A-rate was constant across the sampling sequences in the sampling paradigm, which was consistent with previous findings (e.g., Hau et al., 2008) but was contradicted by Gonzalez and Dutt (2011).

Another assumption in analyzing the decision-from-experience data that neglects individual differences is associated with the number of samples a decision-maker draws before making a final action. Many models have a component for the number of samples, such as the parameter $k$ (for the number of samples drawn) in primed sampler models (Erev et al., 2008b, 2010), and the stopping rule in the IBLT model used to define the number of samples as a random draw from a stochastic distribution (Gonzalez & Dutt, 2011). However, the number of samples is assumed to be equal for all individuals on all decision problems. Moreover, the parameters associated with the number of samples draw are generally estimated from the aggregated data. A quick non-parametric one-way
ANOVA showed that the assumption was not very delicate. For the decision-from-experience sampling data set in the Technion Prediction Tournament, a Kruskal-Wallis rank sum test showed that the number of samples differed by individual ($\chi^2 (39) = 734.50, p < .001$), but not by problems ($\chi^2 (59) = 28.47, p > .05$). The results suggest that individual variability was more salient than item variability with regard to the number of samples drawn. Completely pooling this information can conceal interesting data patterns. From the perspectives of A-rate and the number of samples, it seems that individual differences in the behavioral pattern do exist in decision-from-experience data and are critical in modeling the decision-making processes.

In conclusion and in view of the importance of proper modeling of individual differences, the hierarchical Bayesian modeling approach seems to offer a suitable alternative to integrate the critical model assumptions in accounting for the decision-from-experience experiments and in further exploring the nature of the description-experience gap. The next chapter introduces the hierarchical Bayesian modeling framework with an emphasis on explaining how hierarchical Bayesian modeling can benefit modeling the description-based and experience-based experiments.
3.1 Hierarchical Bayesian modeling

It is commonly accepted that the topic of individual differences is more of a concern to fields such as psychological testing, personality, clinical assessment, etc. In cognitive modeling, however, individual variance traditionally has been ignored because it used data that have been averaged or aggregated across participants (Navarro, Griffiths, Steyvers, & Lee, 2006). The practice of cognitive modeling reveals that when the performance of participants has genuine differences, or some assumptions in averaging data do not hold, then averaging data would lead to a misleading bias (Estes, 1956, Myung, Kim, & Pitt, 2000). Hence, individual differences should be addressed in cognitive modeling as well.

Many approaches have been used in cognitive modeling to account for individual differences in cognitive processes. Under certain circumstances, it is possible to obtain parameter estimates for individuals when the number of data points is sufficiently large. The parameter estimates then can be compared at individual and group levels (e.g., Ahn et al., 2008; Estes & Maddox, 2005; Ratcliff & McKoon, 2008). Structural Equation Modeling (SEM), one of the most popular statistical tools in social science mainly used
for psychological assessment, may be employed to relate the parameter estimates of
cognitive models and scores from psychological tests or inventories (e.g., Ratcliff,
Thapar, & McKoon, 2010). Bayesian methods, and hierarchical Bayesian modeling in
particular, also has become familiar to many cognitive scientists (Kruschke, 2011; Lee,
2011; Tenenbaum, Griffiths, & Kemp, 2006).

Hierarchical Bayesian modeling is a sophisticated way of representing and
modeling individual differences within models of cognitive processes, and has received
much attention in recent years (e.g., Lee, 2008, 2011; Shiffrin et al., 2008). The central
innovation of hierarchical Bayesian modeling is to provide an explicit model for the
kinds of individual differences that might be present in the data. One general approach is
to supplement a cognitive model with an individual differences component that describes
how cognitive parameters can vary across people (Navarro et al., 2006). Using
sufficiently flexible individual differences models, researchers can learn both the
similarities and differences between and among people.

Hierarchical Bayesian modeling offers many advantages in model fitting
compared to the non-hierarchical, individual analysis approach. To give a concrete
example, Rouder, Lu, Speckman, Sun, and Jiang (2005) simulated the Weibull distributed
response time data and examined the performance of various estimation methods in
experimental psychology. The comparison of root mean square error (RMSE) shows that
the hierarchical Bayesian method provides more accurate parameter estimates than
nonhierarchical estimates (e.g., maximum likelihood, quantile maximum likelihood, and
quantile least squares), especially when the number of observations per participant was
small. Nilsson, Rieskamp, and Wagenmakers (2011) and Wetzels, Vandekerckhove,
Tuerlinckx, and Wagenmakers (2010) also show that estimating decision-making models via hierarchical Bayesian method helps avoid boundary estimates and provides more informative inferences.

Whereas individual differences offer a clear justification for the hierarchical Bayesian modeling approach, the usage of hierarchical Bayesian modeling can be much broader. Rouder and Lu (2005) pointed out that, in cognitive experiments, unmodeled variability often yields asymptotic bias and distorted parameter estimates, in particular, for nonlinear models. Hierarchical Bayesian modeling is also used to model participant variability, item variability, and measurement error simultaneously, and hence obtain more accurate parameter estimates in nonlinear models. From a theoretical perspective, Lee (2011) projected some possible uses of hierarchical Bayesian modeling that would better associate cognitive modeling with theoretical development and would contribute to unifying disparate models. Given these advantages of hierarchical Bayesian modeling, it is not surprising to see that the popularity of hierarchical Bayesian modeling has increased in recent years in various areas of cognitive psychology, such as: recognition memory and signal detection paradigm (e.g., Pooley, Lee, & Shankle, 2011; Pratte & Rouder, 2011; Rouder & Lu, 2005); categorical learning (e.g., Lee, 2006; Vanpaemel, 2011); cognitive development (e.g., Scheibeheenne, Rieskamp, & Wagenmakers, 2013; Tenenbaum et al., 2006); confidence judgment (e.g., Merkle, Smithson, & Verkuilen, 2011); risky choice (Nilsson et al., 2011); and dynamic decision-making (e.g., Lee, 2008; Steyvers, Lee, & Wagenmakers, 2009; Wetzels et al., 2010).

On the other hand, there have been some technical challenges with which to deal in the application of hierarchical Bayesian models for cognitive modeling. Implementing
hierarchical Bayesian models typically require tractable likelihood functions and nontrivial effort in equation derivation or programming. Computational load is also of concern because sophisticated numerical methods, such as MCMC simulation, commonly are applied in parameter estimation. Moreover, the method remains under development and has similar problems and critiques usually associated with Bayesian methods (Rouder & Lu, 2005). Although it functions better than other estimation methods when observations are few, the advantage of the hierarchical model is a function of sample size and works better for larger sample size. Other problems also exist, such as over-shrinkage, high autocorrelations, and the dependence between individual parameter estimates. Despite these implementation or computation challenges, the benefit of hierarchical Bayesian modeling outweighs the obstacles.

Since graphical modeling is one of the most popular ways to express and evaluate hierarchical Bayesian modeling, the following section provides a brief overview of graphical models.

3.1.1 Graphical representation of hierarchical Bayesian modeling

Graphical modeling often is employed to represent complex Bayesian models with a large number of random variables. A graphical model defines a family of probability distributions in terms of a graph containing directed or undirected nodes (Jordan, 2004). Each node represents an observed variable or a latent parameter, and nodes are interrelated through deterministic or stochastic dependencies. As such, a graph offers an intuitive and convenient way to capture the hierarchical structures and dependencies between parameters and observed variables in a complex model.
Hierarchical structures in a graphical model usually build upward from the bottom. To demonstrate an example, consider the simple three-layer structure as illustrated in Figure 2. The smallest unit of observations usually consists of the bottom layer, such as the response to a certain item from an individual participant, denoted as $R_i$. The response is determined by an individual-level parameter $\theta_i$ consisting of the middle layer, and the individual-level parameter is a random sample from a group-level distribution with the two parameters $\alpha_{\theta}$ and $\beta_{\theta}$ consisting of the top layer. Dependencies between these nodes are expressed in terms of directed arrows pointing from one layer to another. In the example of Figure 2, the observed response $R_i$ depends on $\theta_i$, so $R_i$ is a child node and $\theta_i$ is its parent node. A node can be child and parent simultaneously, especially in hierarchical models, e.g., $\theta_i$ is itself a child node determined by $\alpha_{\theta}$ and $\beta_{\theta}$.

![Figure 2. A simple hierarchical model in terms of a graphical model.](image)

Graphical models use visual properties of a node to indicate its statistical characteristics. Three characteristics are commonly used following the convention of Lee and Wagenmakers (in press). Firstly, known quantities such as the observed variables and study designs are represented by nodes shaded in gray, and model parameters or other
latent intermediate variables are represented by non-shaded nodes. Secondly, continuous variables are represented by circular nodes, whereas discrete variables are represented by square nodes. Lastly, nodes that are determined completely by the values of their parent nodes usually have double borders, and other stochastic nodes have only single borders. In the present study, for the sake of simplicity, the last one is omitted, and all nodes have single borders. Whether a node is stochastic or deterministic, it can be judged from model descriptions.

One of the disadvantages of graphical representation of a hierarchical Bayesian model is the absence of information about statistical distributions and functional forms in the presented model. Therefore, additional descriptions are required for model priors, structural assumptions among variables and parameters, and statistical distributions of parameters. Assuming that the model in Figure 2 accounts for a binary choice data where $R_i$ is the choice from a Bernoulli trial, $R_i$ can be determined by a latent choice probability $\theta_i$, which represents individual $i$’s probability of endorsing an item to be $R = 1$. Moreover, an individual’s choice probability $\theta_i$ may be beta distributed at the group level with the parameters $\alpha_{\theta}$ and $\beta_{\theta}$. Assuming Jeffreys’ prior, the statistical model definition accompanying the graphical model in Figure 2 is written as:

\[
\begin{align*}
R_i &\sim \text{Bernoulli}(\theta_i) \quad \text{(likelihood)} \\
\theta_i &\sim \text{Beta}(\alpha_{\theta}, \beta_{\theta}) \quad \text{(prior)} \\
\alpha_{\theta}, \beta_{\theta} &\sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right) \quad \text{(hyper-prior)}
\end{align*}
\]
In short, a complete description of a hierarchical Bayesian model should include the statistical definition and parameter priors as above, in addition to the graphical representation.

3.1.2 Estimation of hierarchical Bayesian models

As hinted in its name, the estimation of hierarchical Bayesian models is accomplished using Bayesian statistics. When estimating hierarchical Bayesian models, the posterior distribution of model parameters and the predictive distribution of response variables are typically of interest for inference purposes. Except for very few conjugated models, however, analytical solution is not plausible and simulation methods must be used. In this section, three estimation methods are briefly introduced for the toy model in Figure 2. These methods include the analytical derivation of posterior distribution, conditional simulation, and the Markov chain Monte Carlo (MCMC) algorithm using the Gibbs sampler.

Analytical solutions are usually available only when the model is relatively simple and assumes conjugate priors. That is, the prior and posterior distribution of parameters fall in the same family of distributions (Gelman, Carlin, Stern, & Rubin, 2004, p. 40). For the toy model, the beta prior distribution is a conjugate family for the binomial (and Bernoulli) likelihood. If \( \alpha_0 \) and \( \beta_0 \) have constant values \( \alpha_0 \) and \( \beta_0 \), the posterior distribution for \( \theta \) may be directly computed as:

\[
\theta_i | R_i \sim \text{Beta}(\alpha_0 + R_i, \beta_0 + 1 - R_i) \quad \text{(posterior)}
\]
When αθ and βθ are not fixed values, however, the expression of posterior distribution for θ|R becomes intractable. Conditional simulation is a relatively simple simulation method and is applicable when the conditional distribution of θ|(αθ, βθ, R) and the marginal distribution of (αθ, βθ)|R is known. Table 2 shows the pseudo code that is used to obtain posterior samples for θ.

When the marginal distribution of (αθ, βθ)|R is unknown but the conditional distributions of θ|(αθ, βθ, R) and (αθ, βθ)|(θ, R) are known, the posterior samples for θ can still be obtained using the MCMC algorithm. The pseudo code in Table 3 depicts the procedure of an MCMC algorithm called the Gibbs sampler (a.k.a. alternating conditional sampling, Gelman, Carlin, Stern, & Rubin, 2004, p. 287). In brief, the MCMC algorithm draws the posterior samples by forming a Markov chain in which the stationary distribution converges to the joint posterior distribution of (αθ, βθ)|R, i.e., the posterior distribution of the hyper-parameters.

As the number of parameters and the complexity of link functions in hierarchical Bayesian modeling increases, it becomes more difficult to determine the conditional and marginal distributions and therefore, to implement appropriate simulation procedures. Also, with different specification of distributions, more complex Monte Carlo algorithms may be required to obtain the posterior distributions. Examples of other algorithms include the Metropolis-Hasting algorithm (Chib & Greenberg, 1995), the slice sampling (Neal, 2003), the hybrid Monte Carlo (see Morey, 2011), and the Hamiltonian Monte Carlo (e.g., Hoffman & Gelman, 2011). To help researchers focus more intently on model development, there are software packages developed to assist the implementation of HBM. Two major packages for this purpose are BUGS (Bayesian inference Using Gibbs
Sampling, Lunn, Spiegelhalter, Thomas, & Best, 2009) and JAGS (Just Another Gibbs Sampler, Plummer, 2003). BUGS and JAGS both are developed for analyzing graphical models and performing Bayesian inference using the Gibbs sampler and other MCMC algorithms. Another software package called Stan (Stan Development Team, 2013) recently has been released, which relies on the Hamiltonian Monte Carlo methods. For additional technical details and application examples in psychology, the reader is directed to Bayesian books and journals of psychological studies in Gelman et al. (2004), Kruschke (2011), Lee & Wagenmakers (in press), and the special issue of Journal of Mathematical Psychology on hierarchical Bayesian models (Lee, 2011).

Table 2. Pseudo code for obtaining the posterior samples for θ using conditional simulation.

```
for (k in 1:K) {
    sample alpha.post[k] from p(α|θ)
    sample beta.post[k] from p(β|θ)
    sample theta.post[k, 1:N] from p(θ|α, β, R)
}
```

Table 3. Pseudo code for obtaining the posterior samples for θ using Gibbs sampler.

```
- Choose start values to initiate the Markov chain
  set theta.post[1:N] = theta.0[1:N]
  set alpha.post = alpha.0, beta.post = beta.0

- Draw samples iteratively from the conditional distributions
  for (k in 1:K) {
    sample theta.post[1:N] from p(θ|α=alpha.post, β=beta.post, R)
    set theta.samples[k, 1:N] = theta.post[1:N]
    sample alpha.post and beta.post from p(α, β|θ=theta.post[1:N], R)
    set alpha.samples[k] = alpha.post, beta.samples[k] = beta.post
  }
```
3.1.3 Model comparison

Model comparison is also of importance to the application of hierarchical Bayesian modeling, although well-known measures of model selection may not be directly applicable. For instance, it is not straightforward to determine the number of parameters when hierarchical structure is involved; hence, AIC and BIC may not be the most appropriate model selection measures to use in such situations. Instead of dwelling on the summary and comparison of model selection measures, however, the present project adopts a specific measure of the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002) and validates its use in comparing the hierarchical generalization models. DIC is defined as:

\[
\text{DIC} = \bar{D} + p_D \tag{7}
\]

\[
\bar{D} = -2 \log p(Y|\theta) \tag{8}
\]

\[
p_D = E_{\theta|Y}[-2 \log p(Y|\theta) + 2 \log p(Y|\bar{\theta}(Y)) \tag{9}
\]

where \(\bar{D}\) is the posterior mean deviance which represents model fit, and \(p_D\) is the effective number of parameters which represents model complexity. The conventional decision criterion for DIC is that a model with a smaller DIC value tends to generalize better.

Following Spiegelhalter et al. (2002) and Plummer (2008), \(\bar{D}\) is defined in terms of the logarithmic loss function:

\[
\bar{D} = -2 \int p(\theta|Y) \log[p(Y|\theta)]d\theta \tag{10}
\]
As shown in Equation 9, the definition for \( p_D \) is the difference between the posterior mean deviance and deviance of posterior means. The operational definition and estimation of \( p_D \) has been discussed widely but a complete justification is yet to develop (e.g., Celeux, Forbes, Robert, & Titterington, 2006; Gelman et al., 2004; Plummer, 2008; Spielgelhalter et al., 2002). The present study utilizes the practical definition suggested in the appended discussion of Spielgelhalter et al. (2002) and justified in Plummer (2008). That is, \( p_D \) is estimated as the expected value of \( I(\theta^0, \theta^1) \), defined as:

\[
p_D = E_{\theta|Y}[I(\theta^0, \theta^1)] = E_{\theta|Y} \left\{ E_{Y_{rep}|\theta^0} \left[ \log \frac{p(Y_{rep}|\theta^0)}{p(Y_{rep}|\theta^1)} \right] \right\} \tag{11}
\]

where \( \theta^0 \) and \( \theta^1 \) are two different values of \( \theta \), \( Y_{rep} \) is the replicated values of dependent variables, and \( I(\theta^0, \theta^1) \) represents the Kullback-Leibler information divergence between the predictive distributions \( p(Y_{rep}|\theta) \) at \( \theta^0 \) and \( \theta^1 \). In practice, the value of \( p_D \) is estimated via MCMC sampling by using two parallel Markov chains and taking the sample average of \( \log \frac{p(Y_{rep0}|\theta^0)}{p(Y_{rep1}|\theta^1)} \).

Use of DIC has been reported in psychological studies (e.g., Myung, Karabatsos, & Iverson, 2005; Pratte & Rouder, 2011). DIC can be viewed as a Bayesian analog to AIC, but is known to violate the reparameterization invariance rule (Myung, Karabatsos, & Iverson, 2008). Reparameterization invariance refers to the property of a statistical index that the value of the index does not change when the functional form of a model is
rewritten to an equivalent form with a different set of parameters. The violation of the reparameterization invariance suggests that the interpretation of DIC results should be proceeded with caution. The choice of model comparison methods is revisited in the general discussion.

3.2 Hierarchical Bayesian modeling adaption of CPT

Given that hierarchical Bayesian modeling can account for individual differences in a natural manner, it is a well-suited tool to model the risky choice data observed in not only the decision-from-description paradigm but also the decision-from-experience paradigms. Hierarchical Bayesian modeling can be further extended and used in the hierarchical generalization models, providing a unified approach for modeling common decision-making processes that generate across different tasks. This section discusses how to implement hierarchical Bayesian modeling to model risky choice data in the context of the decision-from-description paradigm.

The mainstream decision-making models for risky choice data are built under the framework of Cumulative Prospect Theory, or CPT, (Luce & Fishburn, 1991; Tversky & Kahneman, 1992). In CPT, each prospect is denoted as \( g = (p_1, r_1; \ldots; p_n, r_n) \), where \( p_i \) is the chance of receiving outcome \( r_i \). For the gamble setup in the present study, the two-outcome prospects are simplified to \( g_r = (p_h, r_h; 1-p_h, r_l) \) and \( g_s = r_m \). The way CPT accounts for the risky choice process is to evaluate the probability and outcome associated with the prospects and assign a subjective value to each prospect by combining all the weights placed on probability and outcome information.
The experimental data of interest are collected in binary-choice gamble games. Specifically, in this study gambles with two prospects are considered. One is a safe prospect with a fixed payoff schedule, and the other is a risky prospect with a probabilistic payoff distribution. The safe prospect always yields a fixed outcome $r_m$. The risky prospect has the probability of $p_h$ to yield an outcome $r_h$, and the probability of $(1-p_h)$ to yield an outcome $r_l$. There are four design variables: payoffs $r_h$, $r_l$, and $r_m$, and the probability $p_h$. The payoff values are subject to the constraint of $r_h > r_m > r_l$ to avoid a dominant strategy, e.g., the safe prospect would be chosen all the time if it always gave the largest reward in that $r_m > r_h > r_l$.

For the aforementioned gamble setting, the subjective values (a.k.a. utilities) assigned to the risky and safe prospects are calculated according to CPT as:

$$SV(g_r) = w(p_h)v(r_h) + w(1-p_h)v(r_l)$$  \hspace{1cm} (12)

$$SV(g_s) = w(p=1)v(r_m)$$  \hspace{1cm} (13)

where $v(x_i)$ is a monotonic value function and $w(p_i)$ is a monotonic probability weighting function constrained by $0 \leq w(p_i) \leq 1$.

Tversky and Kahneman (1992) originally proposed Equation 14 and Equation 15 for the value function and the probability weighting function, respectively. Equation 14 is a power function which takes different forms, depending on whether the gamble outcomes are positive (gains) or negative (losses). $\alpha$, $\beta$, and $\lambda$ are free parameters, with $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, and $\lambda$ as a scaling parameter. Previous studies supporting the prospect theory typically characterize the probability weighting function as having an
inverse-S shape (Gonzalez & Wu, 1999; Tversky & Kahneman, 1992), and Equation 15 is proposed to reflect such a shape. $\gamma$ and $\delta$ are free parameters with $0 \leq \gamma \leq 1$, and $0 \leq \delta \leq 1$. $\gamma$ and $\delta$ may take different values depending on whether the gamble outcomes are positive or negative, as follows:

$$v(r_j) = \begin{cases} r_j^\alpha & \text{if } r_j \geq 0 \\ -\lambda(|r_j|) & \text{if } r_j < 0 \end{cases}$$ (14)

$$w(p_j) = \begin{cases} p_j^\gamma & \text{if } r_j \geq 0 \\ \frac{p_j^\gamma}{(p_j^\gamma+(1-p_j)^\gamma)^\gamma} & \text{if } r_j < 0 \end{cases}$$ (15)

For ease of explanation, the hierarchical Bayesian modeling adaptation of CPT models makes use of gambles with positive outcomes only and is illustrated in Figure 3. Hence, only the top lines of the value function (Equation 14) and the probability weighting function (Equation 15) are used, referred to as $v^+(r)$ and $w^+(p)$. Assume that there are $N$ persons and $N_j$ gambles in a data set. $D_{ij}$ denotes the $i^{th}$ person’s choice in the $j^{th}$ gamble, $D_{ij} = 1$ if the risky prospect is chosen, and $D_{ij} = 0$ if the safe prospect is chosen. The gamble design variables are denoted as $x = \{p, r_l, r_m\}$. $D_{ij}$ and $x_j$ are known quantities and are shaded in gray. Only $D_{ij}$ is a discrete variable represented by a square node.
As shown in Figure 3, the value $D_{ij}$ is a random draw from the Bernoulli trial and with a latent parameter $\theta_{ij}$, which is the probability of endorsing the risky prospects for the $i^{th}$ person in the $j^{th}$ gamble. $\theta_{ij}$ can be stochastically or deterministically determined under the CPT framework. The implementation in the present study follows the exponential Luce’s choice rule (Nilsson et al., 2011; Stott, 2006), where the probability of choosing the risky prospect over the safe prospect is given by

$$\theta_{ij} = \frac{1}{1 + \exp(\varphi(SV(g_s) - SV(g_r)))} \quad \text{(link function)}$$

$$D_{ij} \sim \text{Bernoulli}(\theta_{ij}) \quad \text{(likelihood)}$$
where $\phi$ is a tuning parameter or a free parameter depending on different modeling assumptions, and $SV(g_s)$ and $SV(g_r)$ are the subjective values of the safe and risky prospects defined in Equations 12 and 13.

The individual level parameter $\alpha_i$ for the value function and $\gamma_i$ for the probability weighting function determines its value through $v^+ (r)$ and $w^+ (p)$, defined in Equations 14 and 15. These individual parameters also have their parent nodes, or formally, follow a hyper-distribution with the group level hyper-parameters. Because $\alpha_i$ and $\gamma_i$ are bounded between 0 and 1, intermediate individual parameters $\alpha_i^\phi$ and $\gamma_i^\phi$ are first sampled. Then a probit transformation is used to map the intermediate parameters onto the scale of $[0, 1]$, following Nilsson et al. (2011) and Rouder and Lu (2005).

$$\alpha_i = \Phi(\alpha_i^\phi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha_i^\phi} e^{-t^2/2} dt \quad \text{(probit transformation)}$$

$$\alpha_i^\phi \sim N(\mu_\alpha^\phi, \sigma_\alpha^\phi) \text{ and } \gamma_i^\phi \sim N(\mu_\gamma^\phi, \sigma_\gamma^\phi) \quad \text{(parameter distributions)}$$

Finally, the graphical model requires the specification of priors for hyper-parameters $\mu_\alpha^\phi, \sigma_\alpha^\phi, \mu_\gamma^\phi, \text{ and } \sigma_\gamma^\phi$, such as,

$$\mu_\alpha^\phi, \mu_\gamma^\phi \sim \text{Beta}(1,1) \quad \text{(hyper-prior)}$$

$$\sigma_\alpha^\phi, \sigma_\gamma^\phi \sim \text{Inverse Gamma}(0.001, .001) \quad \text{(hyper-prior)}$$

Another item to notice in Figure 3 is the two plates placed on the nodes. The plate associated with $i = 1 \ldots N$ persons indicates that the contained nodes $D_{ij}, \theta_{ij}, \alpha_i, \text{ and } \gamma_i$ are
variables or parameters related with individual persons. The plate associated with \( j = 1 \ldots N \) indicates that the contained nodes \( D_{ij}, \theta_{ij}, \) and \( x_j \) are related with each gamble game. The joint set of the two plates signifies that the relevant nodes \( D_{ij} \) and \( \theta_{ij} \) vary for each person in each gamble game. All the dependencies and statistical structures, taken together, build up the hierarchical Bayesian modeling account for the CPT framework.

3.3 Implementation of hierarchical generalization models

This section discusses an implementation of hierarchical generalization models in the hierarchical Bayesian modeling framework and discusses how to compare models within the framework of hierarchical generalization models. In particular, the models that assume invariant and partially invariant functional forms are of central interest.

Model comparison is demonstrated for carefully chosen subsets of candidate hierarchical generalization models. Experimental studies have been interested in fitting parameterized functional forms to the CPT (Neilson & Stowe, 2002). Many functional forms have been suggested and tested for the probability weighting and value functions of CPT (for reviews and comparisons, see Gonzalez & Wu, 1999; Neilson & Stowe, 2002; Stott, 2006). The present study follows the common practice of comparing CPT models and use gambles with positive rewards only. Moreover, there has been consensus on value function in that power value function generally outperforms other value functions (see Stott, 2006, for example), whereas the values and interpretations of power value function attract more interest. Hence, the current study will start with holding the functional form of value function to be constant and examining the probability weighting functions.
In addition to the original Tversky and Kahneman (1992) probability weighting function (referred to as TK hereafter), Prelec (1998) axiomatically derived a family of weighting functions, often referred to as the compound-invariance family in Luce (2001). With regard to the positive rewarded gambles, the one-parameter version Prelec probability weighting function is:

\[ w^+(p_j) = \exp\left(-(-\ln p_j)^\gamma\right) \]  

(16)

where \( \gamma \) is a free parameter with \( 0 \leq \gamma \leq 1 \). Equation 16 is usually referred to as Prelec-I in the literature and has a similar inverse-S shape as the original TK weighting function.

The other two functional forms for probability weighting functions considered in the current study are power weighting function and linear weighting function, each with one free parameter. Specifically, the power weighting function is defined as: \( w^+(p_j) = p_j^\gamma \), with a free parameter \( \gamma \) (0.1 ≤ \( \gamma \) ≤ 10). The linear weighting function is defined as: \( w^+(p_j) = w \cdot p_j \), with a free parameter \( w \) (0.1 ≤ \( w \) ≤ 10). The shapes of the four probability weighting functions under consideration are shown in Figure 4. The parameters for the functions are: \( \gamma = 0.71 \) for TK, \( \gamma = 0.74 \) for Prelec-I, \( \gamma = 0.84 \) for power, and \( w = 1.2 \) for linear. These are some examples of possible parameter values in the literature (Stott, 2006).
Figure 4. Representative shapes of four probability weighting functions.

(The legend and data generating parameters for the functions are as follows. Solid curve: TK function, $\gamma = 0.71$; dashed curve: Prelec I function, $\gamma = 0.74$; dotted curve: power function, $\gamma = 0.84$; dotted dashed line: linear function, $w = 1.2$.)

Thus far, the hierarchical Bayesian modeling adaptation of CPT models was discussed. Now, the model illustrated in Figure 4 is further extended to simultaneously model data observed from the two gamble paradigms, the decision-from-description (DFD) paradigm and the sampling paradigm of decision-from-experience (DFE-Sampling or DFE-S). Assume that there are $N$ persons and $N_j$ gambles for each task. $D_{ij1}$ denotes data collected from the decision-from-description paradigm and $D_{ij2}$ denotes data collected from the decision-from-experience, sampling paradigm.

Shown in Figure 5 is a variant of the pFIP1 model that assumes partially generalizable structures across the two tasks. That is, the two tasks share the same structural form of the power value function $v^+(r) = r^\alpha, 0 < \alpha < 1$. The individual level
parameters for the value function, $\alpha_1$ and $\alpha_2$, share the same hyper-parameters $\mu_\alpha$ and $\sigma_\alpha$.

As for the probability weighting function, DFD (Task 1) assumes TK probability weighting function, with the individual parameters $\gamma_{i1}$ and hyper-parameters $\mu_{\gamma1}$ and $\sigma_{\gamma2}$. DFE-S (Task 2) assumes the linear probability weighting function, with the individual parameters $w_{i1}$ and hyper-parameters $\mu_w1$ and $\sigma_w2$. The pFIPi model is considered a partially generalizable model because the underlying value function holds invariant across DFD and DFE-S tasks. What has been learned from the DFD task about the value function can be applied to the value function in the DFE-S task. On the other hand, the knowledge about probability weighting function gained from the DFD task cannot be applied to the DFE-S task, thus the generalization is only partial. The partially generalizable model shows support for certain generalizable processes while allowing the existence of unique additional processes for different experimental tasks.

In contrast to the partially generalizable model, a fully generalizable FIPi model, shown in Figure 6, assumes the same hyper-structures and hyper-parameters for both DFD and DFE-S tasks. For FIPi models, the knowledge about the decision-making process gained from the DFD task can be completely applied to the DFE-S task. Fully generalizable models are desirable in the ideal cases, suggesting that the cognitive decision-making processes underlying DFD and DFE-S tasks are identical.

Because the fully generalizable and the partially generalizable hierarchical models can translate the hypothesis for generalization into testable model assumptions, the next steps in building up the hierarchical generalization models are to: 1) establish a procedure to choose a subset of models; 2) compare the chosen models; 3) pick an appropriate one from the competing pool; and 4) determine which underlying processes are generalizable
across tasks. The specific procedure of comparing hierarchical generalization models is discussed and implemented in the simulation study.

Figure 5. Graphical scheme of a partially generalizable p-FIPI model.

Figure 6. Graphical scheme of a fully generalizable FIPI model.
Chapter 4: Part I – Simulation Study

The purpose of simulation studies in this chapter is two-fold. Firstly, the proposed hierarchical Bayesian modeling approach to account for the cross-task binary choice data are examined to check if it can correctly recover the model structures and parameters that generate data. Secondly, the procedure of comparing hierarchical generalization models is investigated to make sure it can correctly identify the generalizable model components across tasks.

4.1 Study 1: Parameter Recovery Study

Study 1 focuses on the parameter recovery ability of hierarchical Bayesian version of models under the Cumulative Prospective Theory (CPT) framework. Binary choice data for the decision-from-description and decision-from-experience experiments are generated from the models illustrated in Figure 5 and Figure 6, and then fitted to each model, respectively. The choice data are generated for 100 subjects, based on the 20 gambles with positive outcomes from Erev et al. (2010).

The data-generating hyper-parameters are chosen from the best-fitting parameters reported in previous studies (Wu & Gonzalez, 1996). In generating the data, 100 sets of individual parameters for each task are drawn first from the hyper-parameter distributions. Next, $\theta_{ijk}$ ($i = 1 \ldots 100, j = 1 \ldots 20, k = 1, 2$) are calculated from the
parameters and gamble variables, and the outcome of each binary choice is determined by the following exponential Luce’s rule (Nilsson et al., 2011; Stott, 2006), where the probability of choosing the risky prospect over the safe prospect is determined by:

$$\theta(SV(g_r), SV(g_s)) = \frac{1}{1 + \exp(-\varepsilon (SV(g_r) - SV(g_s)))}$$ (17)

where $\varepsilon = 3$ is set to a constant for simplification purposes.

The models are fitted in the software package JAGS 3.3.0 (Plummer, 2003) called by the package of rjags (Plummer, 2012) from R 2.15.2 (R Core Team, 2012). The main reason for choosing JAGS is to facilitate parallel computing and ensure that each Markov chain starts with a different random seed. Model fits and parameter estimates from JAGS have been compared to OpenBUGS (Lunn et al., 2009) results called by both packages of BRugs (Thomas, O’Hara, Ligges, & Sturtz, 2006) and R2OpenBUGS (Sturtz, Ligges, & Gelman, 2005) and have been proved to be similar. For parameter recovery, model fitting in JAGS is computed by running two parallel chains for 500,000 iterations, with discarding the first 5,000 samples, and the thinning interval is set at 100.

4.1.1 Fully generalizable model

The fully generalizable model for the binary gamble games in the decision-from-description and decision-from-experience experiments assumes the same functional forms and hyper-parameter distributions for both tasks, namely, the power value function in Equation 14 and the TK probability weighting function in Equation 15. The data-generating hyper-parameters are $\mu_\alpha = .48$ for the power value function, $\mu_\gamma = .71$ for the
TK probability weighting function, and $\sigma_{\mu_\alpha} = \sigma_{\mu_\gamma} = .01$. When fitting the model, the means of hyper-parameter distribution were monitored and a total of 10,000 posterior samples were obtained for each parameter.

Convergence at the end of simulation was examined by various diagnostics. Trace plots showed that both chains converge to similar values. Numerical diagnostic methods (Cowles & Carlin, 1996) also supported convergence with $\bar{R} = 1$ (Gelman & Rubin, 1992, Gelman et al., 2004), and passed diagnosis tests of both the Geweke test and the Heidelberger and Welch test. Table 4 lists the summary statistics of the posterior samples of the hyper-parameters. Figure 7 illustrates the estimated density of these posterior samples and their 95% credible intervals. The top panel shows the estimated density of $\mu_\alpha$ and $\sigma_\alpha^p$, and the bottom panel shows the estimated density of $\mu_\gamma$ and $\sigma_\gamma^p$. The black dotted lines indicate the data-generating parameter values. The gray dotted lines indicate the positions of 95% credible intervals.

The results from Table 4 and Figure 7 both show that the 95% credible intervals have captured the data-generating values of the hyper-means, suggesting that the functional forms and parameter values are correctly recovered in the Hierarchical Bayesian model. Parameter recovery of the fully generalizable model also shows results comparable to those of Nilsson et al. (2011) in which the hierarchical Bayesian modeling approach was first applied to the CPT framework. In short, the results support the soundness of the hierarchical Bayesian modeling approach in recovering the structure of the fully generalizable model.
Table 4. Summary statistics for means of hyper-parameter distributions in the fully generalizable model.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>Credible Intervals 2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_\alpha )</td>
<td>.462</td>
<td>.463</td>
<td>.031</td>
<td>3.07E-4</td>
<td>.400</td>
</tr>
<tr>
<td>( \sigma_\phi_\alpha )</td>
<td>.230</td>
<td>.235</td>
<td>.113</td>
<td>1.13E-3</td>
<td>.015</td>
</tr>
<tr>
<td>( \mu_\gamma )</td>
<td>.711</td>
<td>.711</td>
<td>.016</td>
<td>1.58E-4</td>
<td>.679</td>
</tr>
<tr>
<td>( \sigma_\phi_\gamma )</td>
<td>.061</td>
<td>.053</td>
<td>.044</td>
<td>4.44E-4</td>
<td>.002</td>
</tr>
</tbody>
</table>

Note: The estimates are based on 10,000 posterior samples from 2 parallel chains. Data-generating values were \( \mu_\alpha = .48 \), \( \mu_\gamma = .71 \), and \( \sigma_\mu_\alpha = \sigma_\mu_\gamma = .01 \).

Figure 7. Distributions of hyper-parameter means estimated from the posterior sample in fitting the fully generalizable model.

(Top left: \( \mu_\alpha \); top right: \( \sigma_\phi_\alpha \); bottom left: \( \mu_\gamma \); bottom right: \( \sigma_\phi_\gamma \). Black dotted lines: data-generating parameters; gray dotted lines: 95% credible intervals.)

4.1.2 Partially generalizable model

For the partially generalizable model in Figure 5, the parameters are \( \mu_\alpha = .48 \) for the common power value function, \( \mu_\gamma_1 = .71 \) for the TK probability weighting function in the decision-from-description task, \( \mu_{w2} = 1.2 \) for the linear probability weighting function in the decision-from-experience experiment, and \( \sigma_\mu_\alpha = \sigma_\mu_\gamma_1 = \sigma_\mu_{w2} = .01 \). In Figure 8, the top panel displays the estimated density of \( \mu_\alpha \) and \( \sigma_\phi_\alpha \), the middle panel reveals the
estimated density of $\mu_{y1}$ and $\sigma_{y1}$, and the bottom panel presents the estimated density of $\mu_{w2}$ and $\sigma_{w2}$. The black dashed lines indicate the data-generating parameter values. The gray dashed lines indicate the positions of 95% credible intervals. Again, the results from Table 5 and Figure 8 both show that the 95% credible intervals have captured the data-generating values, suggesting that the functional forms and parameter values are recovered correctly in the hierarchical generalization model. The proper parameter recovery of partially generalizable models shows that the hierarchical Bayesian modeling approach can simultaneously account for the generalizable component across tasks and the unique component for each task.

Table 5. Summary statistics for means of hyper-parameter distributions in the partially generalizable model.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>Credible Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{\alpha}$</td>
<td>.460</td>
<td>.461</td>
<td>.029</td>
<td>2.92E-04</td>
<td>.402</td>
</tr>
<tr>
<td>$\sigma_{\alpha}$</td>
<td>.061</td>
<td>.051</td>
<td>.047</td>
<td>4.66E-04</td>
<td>.002</td>
</tr>
<tr>
<td>$\mu_{y1}$</td>
<td>.719</td>
<td>.719</td>
<td>.020</td>
<td>2.04E-04</td>
<td>.680</td>
</tr>
<tr>
<td>$\sigma_{y1}$</td>
<td>.085</td>
<td>.075</td>
<td>.060</td>
<td>6.05E-04</td>
<td>.004</td>
</tr>
<tr>
<td>$\mu_{w2}$</td>
<td>1.22</td>
<td>1.19</td>
<td>.015</td>
<td>1.52E-04</td>
<td>1.190</td>
</tr>
<tr>
<td>$\sigma_{w2}$</td>
<td>.016</td>
<td>.014</td>
<td>.011</td>
<td>1.11E-04</td>
<td>.001</td>
</tr>
</tbody>
</table>

Note: The estimates are based on 10,000 posterior samples from two parallel chains. Data-generating values were $\mu_{\alpha} = .48$, $\mu_{y1} = .71$, $\mu_{w2} = 1.2$, and $\sigma_{\mu_{\alpha}} = \sigma_{\mu_{y1}} = \sigma_{\mu_{w2}} = .01$. 
Figure 8. Distributions of hyper-parameter means estimated from the posterior sample in fitting the partially generalizable model.

(Top left: $\mu_\alpha$; top right: $\sigma_\alpha^g$; middle left: $\mu_\gamma_1$; middle right: $\sigma_\gamma_1^g$; bottom left: $\mu_w^2$; bottom right: $\sigma_w^2$. Black dotted lines: data-generating parameters; gray dotted lines: 95% credible intervals.)

4.2 Study 2: Model Recovery Study

Study 1 implemented the hierarchical Bayesian modeling approach in the context of CPT framework, and checked the soundness of the approach. In particular, structural assumptions and parameter values can be recovered through hierarchical Bayesian modeling. An implication of this encouraging result is that with the hierarchical generalization models, it may be possible to test theoretical assumptions in terms of model components (i.e., functional forms and parameter values). This is explored in the next section. Specifically, a model comparison procedure of hierarchical generalization models is constructed to test whether a certain component is generalizable across tasks.
This is performed by turning the target component on and off in the hierarchical generalization models and conducting a model comparison analysis.

Consider again the fully generalizable model in Figure 6 and the partially generalizable model in Figure 5, for instance. The difference between the two models is the structural assumption in terms of functional forms and their related parameters. For the fully generalizable model, the same TK probability weighting function is assumed for both description- and experience-based tasks. For the partially generalizable model, however, the TK probability weighting function is assumed for the description-based task only, but for the experience-based task, the linear weighting function is assumed. The underlying theoretical hypothesis of the partially generalizable model is that subjects would weigh the probability information differently when performing the experience-based task, indicating that the description- and experience-based tasks are performed by weighing the probability information differently. Hence, the probability weighting function does not generalize across experimental settings, suggesting that the probability weighting function may not be a stable cognitive process. On the other hand, the partially generalizable model supports the power value function as a generalizable component across tasks, indicating that the value function is considered a stable cognitive process under this model.

Although the fully generalizable model and partially generalizable model may convey different levels of generalization assumptions, they may both fit a data sample equally well. A rigorous way to test the generalization assumption, then, is to conduct a model comparison analysis with the competing models and determine which one best describes the data. The competing models should include fully generalizable models and
partially generalizable models to enable direct comparisons between functional forms and/or parameters. Therefore, comparison of hierarchical generalization models should contain at least three steps: 1) construct a set of hierarchical generalization models; 2) calculate a model selection measure for each model in the set; and 3) choose an appropriate model, determine the generalizable model component, and interpret in terms of the generalization assumption. A proper model selection measure to combine with hierarchical generalization models is a crucial choice which has an immediate impact on the procedure. Instead of dwelling on the comparison of model selection measures, however, the present study focuses on the use of DIC combined with hierarchical generalization models.

Study 2 is designed to examine the performance of DIC to compare hierarchical generalization models. In particular, Study 2 focuses on evaluating whether the comparison of hierarchical generalization models can successfully identify the data-generating model among a set of competing models. To fully demonstrate the performance of DIC in model comparison, data are generated from each model in a set of competing models and fitted to all candidate models to estimate the values of DIC. The data-generating model is expected to have the smallest DIC value.

The main advantage of hierarchical generalization models over single measures of model selection is its comprehensive consideration of theoretical hypotheses under investigation. With carefully constructed model comparisons, hierarchical generalization models can help identify stable cognitive processes that underlie the experimental tasks of interest. To demonstrate this, rather than considering all possible combinations of candidate functional forms, two chosen sets of models with specific theoretical
implications are considered in Study 2. For simplicity, the same functional form (i.e., the power function in Equation 14) is assumed for both description- and experience-based tasks.

The first set of competing models contains fully generalizable models only. The theoretical implications underlying the fully generalizable models are that the description- and experience-based tasks contain identical cognitive processes in the evaluation of the value and probability information of the prospects. Obviously, this is a strong assumption of generalization. The second set of competing models contains a fully generalizable model and some partially generalizable models that share the same value function in both tasks, but assume different probability weighting functions for the description- and experience-based tasks. As such, the partially generalizable models make a weaker assumption of generalization while still attempting to detect certain generalizable components between the description- and experience-based tasks.

In what follows, simulation studies are conducted with the two sets of competing models described above. The studies are outlined according to the three-step procedure described earlier. The specific choice of competing models are described first, followed by the data generating and model fitting procedures. The simulation results are then analyzed and discussed.

4.2.1 Fully generalizable models

The first set of competing models includes four fully generalizable models listed in Table 6. All models assume the same functional forms and hyper-parameter distributions for data in both decision-from-description (DFD) and decision-from-experience (DFE) tasks. Among the four probability weighting functions, the original TK
and Prelec I functions are both inverse-S shape functions and are very similar to each other. The power and linear functions, on the other hand, are more distinguishable from each other and from the inverse-S shape functions.

Table 6. Competing fully generalizable models.

<table>
<thead>
<tr>
<th>Models</th>
<th>DFD and DFE Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DFD and DFE Tasks</td>
</tr>
<tr>
<td>TK-TK</td>
<td>Value function</td>
</tr>
<tr>
<td></td>
<td>Probability weighting function</td>
</tr>
<tr>
<td>Prelec I-Prelec I</td>
<td>power (α)</td>
</tr>
<tr>
<td></td>
<td>TK (γ)</td>
</tr>
<tr>
<td>Power-Power</td>
<td>Prelec I (γ)</td>
</tr>
<tr>
<td>Linear-Linear</td>
<td>power (α)</td>
</tr>
<tr>
<td></td>
<td>power (γ)</td>
</tr>
<tr>
<td></td>
<td>linear (w)</td>
</tr>
</tbody>
</table>

In the simulation study, data are generated from the four models in turn, and then fit to each model, respectively. The data-generating hyper-parameters are $\mu_\alpha = .48$ for the power value function, $\mu_\gamma = .71$ for the TK function, $\mu_\gamma = .74$ for the Prelec one-parameter function (Prelec I, Prelec, 1998), $\mu_\gamma = .84$ for the power probability weighting function, and $\mu_w = 1.2$ for the linear weighting function. For each simulation, data are generated for 50 and 100 subject with individual parameters distributed with standard deviations ($\sigma_\alpha$, $\sigma_\gamma$, etc.) of .05 and .01, yielding four repetitions. Models are fit in JAGS by running two parallel chains for 100,000 iterations each, with discarding the first 5,000 samples and the thinning interval is set at 1. As a technical note, reducing the thinning interval from 100 to 1 has affected the autocorrelation in posterior samples, but has no visible effect on convergence or on the estimates of $\bar{D}$ and $p_D$. The means of hyper-parameter distributions are monitored, together with $\bar{D}$ and $p_D$. Unless otherwise noted, results are based on simulated data with 100 subjects and parameter standard deviations of .01.
Table 7 shows simulation results for data generated by the fully generalizable model of Model TK-TK, summarized in terms of $\bar{D}$, $p_D$, and DIC values. $\bar{D}$ represents the model fit, whereas $p_D$ represents the model complexity. As displayed in the table, the values of $\bar{D}$ are smallest when the fitted model is the data-generating model. That is, the posterior distribution of the risky choices estimated by Model TK-TK is closer to the simulated data set compared to the three other competing models. Moreover, the value of $p_D$ approximates the effective number of parameters in the hierarchical generalization models. The working definition of $p_D$ in this present study represents a comprehensive evaluation of model parameters and functional form. Larger $p_D$ values generally indicate more complex models, conditional on the data sample. The models considered in Table 7 are all fully generalizable models. Model Prelec I-Prelec I has a highly nonlinear functional form for probability weighting function, and therefore show a relatively large $p_D$ value compared with Model Power-Power and Model Linear-Linear. Although Model TK-TK also has a highly nonlinear probability weighting function, the $p_D$ value is relatively lower because it is the data-generating model for the data sample. In short, the $p_D$ values in Table 7 show a reasonable pattern with respect to its definition, and thus, is working appropriately. The DIC value combines the model fit in $\bar{D}$ and the model complexity in $p_D$ to give a general account for evaluating the models. A smaller DIC value indicates a better model as the description of the embedded structure in the data sample. In Table 7, the data-generating model TK-TK has the smallest DIC value, indicating that the DIC-based model comparison has identified the correct model.

Table 8 lists the DIC values of fitting each competing model to the generated data sets. The data-generating models are listed in rows, whereas the fitted models are
listed in columns. The numbers on the diagonal (bold and underlined) represent the model fit to the "true" models and should be the smallest in each row. The results in Table 8 indicate that the DIC-based model comparison has correctly identified the data-generating model in each of the four cases. As expected, the difference of the DIC values among power function, linear function, and inverse-S shape functions (i.e., TK and Prelec I) are more obvious. The two inverse-S shaped functions are closer to each other in terms of the DIC values. It shows that DIC values are working appropriately to distinguish the functional forms. A natural question to ask is how large a difference should be to in order make a meaningful interpretation. This question will be revisited later.

Table 7. Estimates for DIC related values, and the difference in DIC values.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
<th>DIC difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>5098</td>
<td>20</td>
<td>5118</td>
<td>0</td>
</tr>
<tr>
<td>Prl I-Prl I</td>
<td>5115</td>
<td>39</td>
<td>5154</td>
<td>36</td>
</tr>
<tr>
<td>Pow-Pow</td>
<td>5195</td>
<td>23</td>
<td>5218</td>
<td>100</td>
</tr>
<tr>
<td>Lin-Lin</td>
<td>5173</td>
<td>20</td>
<td>5193</td>
<td>74</td>
</tr>
</tbody>
</table>

Table 8. DIC values estimated from fitting simulated data to the competing fully generalizable models.

<table>
<thead>
<tr>
<th></th>
<th>TK-TK</th>
<th>Prl I-Prl I</th>
<th>Pow-Pow</th>
<th>Lin-Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from TK-TK</td>
<td>5118</td>
<td>5154</td>
<td>5218</td>
<td>5193</td>
</tr>
<tr>
<td>Data from Prl I-Prl I</td>
<td>5314</td>
<td>5308</td>
<td>5369</td>
<td>5378</td>
</tr>
<tr>
<td>Data from Pow-Pow</td>
<td>5534</td>
<td>5535</td>
<td>5361</td>
<td>5439</td>
</tr>
<tr>
<td>Data from Lin-Lin</td>
<td>5548</td>
<td>5550</td>
<td>4084</td>
<td>3619</td>
</tr>
</tbody>
</table>

4.2.2 Partially generalizable models

The second set of competing models includes the fully generalizable model that assumes the TK probability weighting function, and another three partially generalizable
models that share the TK weighting function on one task, but assume a different weighting function on the other task. The four competing models are listed in Table 9. The four models all assume the same functional forms and hyper-parameter distributions for the value function in both description- and experience-based tasks. The underlying theoretical hypothesis of the partially generalizable model is that in the description- and experience-based tasks, subjects would evaluate the gamble outcomes similarly, but would weigh the probability information using different mechanisms.

In the simulation study, data are generated from the four models in turn, and then fit to each model, respectively. The data-generating hyper-parameters are $\mu_\alpha = .48$ for the power value function. For the DFD task, $\mu_{\gamma_1} = .71$ for the TK function all models. For the DFE task, $\mu_{\gamma_1} = .71$ for the TK function, $\mu_{\gamma_2} = .74$ for the Prelec one-parameter function (Prelec I, Prelec, 1998), $\mu_{\gamma_2} = .84$ for the power probability weighting function, and $\mu_{w_2} = 1.2$ for the linear weighting function. Data are generated for 50 and 100 subject with individual parameters distributed with standard deviations of .01, yielding four repetitions for each set of simulations. Unless otherwise noted, results are based on simulations with 100 subjects and parameter standard deviations of .01.

Table 9. Competing partially generalizable models.

<table>
<thead>
<tr>
<th>Models</th>
<th>DFD (Task 1)</th>
<th>DFE (Task 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>TK ($\gamma_1$)</td>
<td>TK ($\gamma_1$)</td>
</tr>
<tr>
<td>TK-Prl I</td>
<td>TK ($\gamma_1$)</td>
<td>Prelec I ($\gamma_2$)</td>
</tr>
<tr>
<td>TK-Pow</td>
<td>TK ($\gamma_1$)</td>
<td>power ($\gamma_2$)</td>
</tr>
<tr>
<td>TK-Lin</td>
<td>TK ($\gamma_1$)</td>
<td>linear ($w_2$)</td>
</tr>
</tbody>
</table>
Table 10 shows results for the data generated by the fully generalizable model of TK-TK and lists the estimates of $\bar{D}$, $p_D$, and DIC values. Similarly, Table 11 shows results for the data generated by one of the partially generalizable models, TK-Power. The fit statistics corresponding to the data-generating models are bold and underlined. The DIC values of fitting the data-generating models remain the lowest compared to the DIC values of fitting the competing models. Some model pairs seem to have a much smaller differences in DIC values. For example, Table 10 shows that, when fitting the simulated data generated from Model TK-TK to Model TK-Prelec I, the estimated DIC value is 5146, which has a difference of 11 compared to the smallest DIC value obtained when fitting the simulated data to its data-generating model TK-TK. The other two competing models, namely TK-Power and TK-Linear, show a difference of 56 and 51 in DIC values, respectively. In such a situation, one may hesitate to make a conclusive decision between TK-TK and TK-Prelec I.

Table 10. Estimates for DIC related values with data generated from Model TK-TK.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
<th>DIC difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>5112</td>
<td>23</td>
<td>5135</td>
<td>0</td>
</tr>
<tr>
<td>TK-Prelec I</td>
<td>5101</td>
<td>45</td>
<td>5146</td>
<td>11</td>
</tr>
<tr>
<td>TK-Power</td>
<td>5156</td>
<td>35</td>
<td>5191</td>
<td>56</td>
</tr>
<tr>
<td>TK-Linear</td>
<td>5163</td>
<td>23</td>
<td>5187</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 11. Estimates for DIC related values with data generated from Model TK-Pow.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
<th>DIC difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>5329</td>
<td>73</td>
<td>5440</td>
<td>199</td>
</tr>
<tr>
<td>TK-Prelec I</td>
<td>5367</td>
<td>25</td>
<td>5353</td>
<td>112</td>
</tr>
<tr>
<td>TK-Power</td>
<td>5219</td>
<td>22</td>
<td>5241</td>
<td>0</td>
</tr>
<tr>
<td>TK-Linear</td>
<td>5257</td>
<td>21</td>
<td>5277</td>
<td>26</td>
</tr>
</tbody>
</table>
Table 12 summarizes the cross-fit of all four competing models in Table 9. That is, data are generated from the four models in turn and are then fit to all the models and the DIC values are estimated. The results show a consistent trend that the DIC values of fitting the data-generating models are the smallest, suggesting that the DIC-based model comparison is correctly identifying the data-generating models under all considered conditions, with the caveat that some model pairs have a much smaller difference in the DIC values.

Cross-fit results are also available for three additional sets of competing models. Each of these additional sets includes a fully generalizable model listed in Table 6, and three corresponding partially generalizable models. To avoid repetition, the simulation results are included in Appendix A. Close examination of the DIC values reveal a similar trend as described above, that the smallest DIC value identified the correct data-generating model all the time, with a few model pairs showing less clear-cut differences in the DIC values. These cases are further analyzed in the following section.

Table 12. DIC values estimated from fitting simulated data to the competing fully and partially generalizable models.

<table>
<thead>
<tr>
<th></th>
<th>TK-TK</th>
<th>TK-Prelec I</th>
<th>TK-Power</th>
<th>TK-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from TK-TK</td>
<td>5135</td>
<td>5146</td>
<td>5191</td>
<td>5187</td>
</tr>
<tr>
<td>Data from TK-Prelec I</td>
<td>3949</td>
<td><strong>3793</strong></td>
<td>3827</td>
<td>3841</td>
</tr>
<tr>
<td>Data from TK-Power</td>
<td>5440</td>
<td>5353</td>
<td><strong>5241</strong></td>
<td>5277</td>
</tr>
<tr>
<td>Data from TK-Linear</td>
<td>5505</td>
<td>5427</td>
<td>4694</td>
<td><strong>4388</strong></td>
</tr>
</tbody>
</table>

4.2.3 Estimation of DIC differences

One major constraint for using DIC is that there is no clear guideline to determine a meaningful threshold when comparing two DIC values. To closely examine the size of
DIC differences, some model pairs with less clear-cut differences in the DIC values are re-examined in order to obtain an empirical distribution of the DIC estimates. To narrow down the size of the replication, the following criteria were applied to choose the model pairs to compare. Above all, the models assuming the Prelec I probability weighting functions in either or both tasks were excluded for two reasons. Firstly, the parameter recovery of Prelec I function is less stable compared to the other three functions. Secondly, the Prelec I and TK functions have a closer relationship in that they both have an inverse-S shape. It would be redundant to consider both functions in a relatively large-scale replication.

For the rest of the models, a rough cutoff of about 50 in the DIC differences was chosen to determine inconclusive decisions. The cutoff value is an arbitrary choice determined by examining all the cross-fitting results in Table 12 and Appendix A. Three model pairs were selected according to the cutoff value for further investigation: TK-TK vs. TK-Linear (when data are generated from TK-TK), TK-Power vs. TK-Linear (when data are generated from TK-Power), and Power-Power vs. Power-Linear (when data are generated from Power-Power).

After settling on the three problematic model pairs, 100 replications of model recovery simulation were performed. In each replication, data were generated from the specified data-generating model, and were fit to the model pair. The data-generating parameters were the same as the previously used parameters for the corresponding functions. Regarding the concern of computational load, data were generated for 20 subjects, and only $\bar{D}$ and $p_0$ were monitored when fitting the hierarchical Bayesian models. Table 13 lists the summary statistics for differences in the estimated DIC values.
When using the rank of DIC as a model selection criterion, that is, when choosing the model with a smaller DIC value regardless how small a difference, the model recovery rates were 97%, 92%, and 93% for the three model pairs, respectively, as illustrated in Figure 9. The model pairs in Figure 9 from left to right are: TK-TK vs. TK-Linear, TK-Power vs. TK-Linear, and Power-Power vs. Power-Linear. The dotted lines in Figure 9 indicate zero DIC difference, and the dashed line indicates the potential cutoff interval using the 5% quantile. The proportion left of the dotted line will be identified incorrectly as the data-generating model.

Several cutoff values were attempted to investigate the potential impact of mis-identification. The first attempt is to use the lower 5% limit as the cutoff criterion. That is, if a difference in the DIC value is larger than the lower 5% limit, the model with a smaller DIC value will be chosen. The dashed lines in Figure 9 indicate the positions of the lower 5% limit at both positive and negative values. The DIC differences between each pair of dashed lines are considered as indistinguishable, whereas the DIC differences outside the lines lead to a conclusive decision. In other words, the instances left of the negative lower 5% limit would be identified incorrectly as data-generating models. The lower 5% limit criterion yielded a mis-identification rate of 1%, 5%, and 5% for the three model pairs, respectively. Other heuristic choices of cutoff criterion are also attempted, such as the cutoff values of 1, 2, and 5. When the cutoff value is set at 5, there are no mis-identified cases, but only 84%, 69%, and 76% of the cases are considered as conclusive decisions, and the rest are indistinguishable.

Setting the cutoff criterion to be larger will help avoid the mis-identification of data-generating model, but also will reduce the model recovery rate by increasing the
number of indecisive instances. This trade-off is not surprising. However, the summary 
statistics in Table 13 and the histograms in Figure 9 also suggest that the prescreening 
cutoff value of 50 may be unnecessarily strict, especially when the sample size is small, 
as the average DIC difference of the three model pairs was around 7 to 11. 
Coincidentally, in two of the three problematic model pairs, the data-generating model 
was the fully generalizable model. Hence, it could be reasonable to penalize less 
generalizable models and choose models with a larger degree of generalization when the DIC differences cannot lead to a clear decision. Adding a penalty term to account for the generalization assumption is also an interesting topic for future research.

Table 13. Summary statistics for DIC differences of the problematic model pairs.

<table>
<thead>
<tr>
<th>Model Pairs</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Error</th>
<th>Quantiles</th>
<th>Minimum DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK vs. TK-Lin</td>
<td>11</td>
<td>12</td>
<td>5.9</td>
<td>0.2</td>
<td>22.0</td>
</tr>
<tr>
<td>TK-Pow vs. TK-Lin</td>
<td>7</td>
<td>7</td>
<td>5.3</td>
<td>-1.8</td>
<td>11.8</td>
</tr>
<tr>
<td>Pow-Pow vs. Pow-Lin</td>
<td>9</td>
<td>9</td>
<td>6.4</td>
<td>-1.4</td>
<td>25.4</td>
</tr>
</tbody>
</table>

Note: The column of Min DIC lists the smallest estimated DIC values in order to give a crude sense of the scale.
Figure 9. Histograms of the DIC difference for the three problematic model pairs.

(Left: TK-TK vs. TK-Linear; middle: TK-Power vs. TK-Linear; right: Power-Power vs. Power-Linear. Dotted lines: 0 difference in DIC values; dashed lines: cutoff criteria using the 5% quantiles of corresponding DIC differences.)
Chapter 5: Part II – Secondary Data Analysis

The second part of the current study presents a secondary data analysis, based on the experimental data previously collected and reported by other researchers. In particular, the secondary data analysis focuses on a comprehensive data set collected in a modeling competition called the Technion Prediction Tournament (TPT, Erev et al., 2010). Applying hierarchical generalization models and model comparison procedures to the data set can help elucidate some of the theoretical question that motivated the TPT.

5.1 Technion Prediction Tournament

Earlier empirical results from the decision-from-experience (a.k.a. experience-based) experiments have exhibited an obviously different trend of bias with the classical decision-from-description (a.k.a. description-based) paradigm of monetary gambles (Barron & Erev, 2003; Hertwig et al., 2004). As the decision-from-experience paradigm attracts research attention, one inevitable shortcoming often criticized is that these studies have a much smaller sample size in terms of number of gamble games and number of trials decision-makers experienced (Fox & Hadar, 2006). At the same time, various quantitative models or heuristics have been developed to account for the decision-from-experience data (e.g., Erev & Barron, 2005; Hau et al., 2008; Yechiam & Busemeyer,
2006). Model comparison methods have been attempted (Hau et al., 2008) but the relatively small sample size remained a problem for model comparison as well.

Under this context, Erev, Ert, and Roth (2008a) organized the Technion Prediction Tournament (TPT). They collected a large and comprehensive data set for three different risky choice paradigms, and invited researchers from all over the world to a prediction competition. The competition data set was collected in two stages. The first part of the data set was collected and published online (Erev et al., 2008a) in April 2008. The second part of the data set was collected months later and was used as a validation data set to evaluate the predictive performance of submitted models. Interested researchers were challenged to submit quantitative models to predict the proportions of risky choice in the validation data set. In other words, the model comparison criterion used in the TPT was the mean squared distance (MSD) between the model predictions and validation data. Fourteen teams of researchers participated in the TPT by submitting models for one or more risky choice paradigms.

Unsurprisingly, the best predictions of the description- and experience-based data sets came from models that were very different from one another. The winning model for the decision-from-description data was a logistic regression model, of which the predictors were the gamble variables, the expected value of the risky prospect, and whether or not the risky prospect had a larger expected value than the safe prospect. The winning model in the sampling paradigm of decision-from-experience data was a model called the ensemble model, which averaged predictions from four different sub-models. The four sub-models considered by the ensemble model included two variants of the natural-mean heuristic, a stochastic CPT model, and a stochastic version of priority
heuristic. The winning model in the repeated-choice paradigm of the decision-from-experience data was a model using the ACT-R architecture, which applied the similarity-based inference and implemented a stochastic softmax retrieval process as the choice rule. Generally speaking, the competition results of the TPT not only revealed a substantial description-experience gap in the empirical data, but also showed a description-experience gap in the models that best predicted the risky choice behavior. Moreover, the models that had a good predictive performance for the description-based data all assumed the processing of probability information. On the other hand, the models for the experience-based data did not include probability as a central component in the decision-making process, and tended to rely on small samples from the experienced outcomes during the experiment. What factors cause the substantial description-experience differences, and how do the decision-making processes differ between description and experience paradigms? These questions remain open after the TPT competition. The secondary data analysis study thus intends to use the hierarchical generalization modeling framework to answer these questions. Some of the models with good predictive performances in the TPT are chosen as the candidate models for the current study, and they are discussed in the following section.

5.2 Models from TPT

This section reviews the candidate models to investigate in the decision-from-description experiment and the sampling paradigm of the decision-from-experience experiment. The primary focus in this study is the generalization of cognitive processes across different paradigms; hence, the dynamic feature in the experience-based
paradigms is not taken into consideration when choosing the candidate models. The chosen models are capable to simultaneously account for the description- and experience-based data. These models are the Cumulative Prospect Theory (CPT) based models and the natural-mean heuristic. The CPT framework is not only the mainstream modeling framework widely used to account for risky choice data, but also the basic component of the many models in the TPT competition. Similarly, the natural-mean heuristic is the basic component of many models used to account for the decision-from-experience data, and is a major component of the winning model in the TPT competition for the sampling paradigm of the experience-based data. Moreover, the natural-mean heuristic is mathematically equivalent to the expected utility and is a special case of the CPT; thus, it may be applied to the description-based data. Other models mentioned in the TPT competition also were considered in the preliminary analysis, but were then excluded for several reasons. For example, the aforementioned winning model of logistic regression did not fit the experience-based sampling data well. Moreover, the hierarchical Bayesian modeling adaption of the logistic regression model was problematic, partly due to the multicolinearity among the predictors. The priority heuristic, which was a component of the winning ensemble model in the TPT sampling data, did not fit the description-based data. The ACT-R model, which uses a dynamic process and won the competition in modeling the repeated-choice data, did not adapt to the description-based data. As a result, the present study only considers the CPT-based models and the natural-mean heuristic.
What follows briefly introduces how these models are implemented in the decision-from-description and decision-from-experience sampling paradigms. The hierarchical Bayesian modeling adaptation of these models also is outlined.

5.2.1 CPT-based models

Despite the fact that earlier experience-based experiments consistently show behavioral patterns that conflict with the patterns commonly observed in the description-based paradigm (Hertwig et al., 2004; Hertwig & Erev, 2009), Fox and Hadar (2006) claim that the differences stem from the larger noise due to sampling error, but the main decision-making mechanism still follows the prospect theory. They proposed the two-stage CPT model to account for the experience-based sampling data. By using the two-stage CPT model, the sampling probabilities experienced during the experiment first were estimated for each problem. Then the utility (i.e., subjective value) for each deck was calculated using the probability weighting function by applying the sampling probabilities. The choice of prospect was based on the utilities.

Figure 10 illustrates the hierarchical Bayesian modeling setup of the two-stage CPT model. There are two major differences compared to the model introduced in Section 3.2. Firstly, because the experienced probabilities and observed outcomes differ for individual participants on each gamble game, the design variables \( x = \{p_i, r_h, r_l, r_m\} \) are no longer constants across subjects. Instead, \( p_{ij} \) and \( r_{ij} \) vary for each subject and are contained in both item and person plates. Secondly, the priors for hyper-parameters \( \mu_a, \sigma_a, \mu_r, \) and \( \sigma_r \) remain unspecified at this point. The hyper-priors will be discussed along with the analysis.
5.2.2 Natural-mean heuristic

The natural-mean heuristic originally was proposed in the reinforcement learning literature for $n$-armed bandit problems (Sutton & Barto, 1998) and was introduced to the experience-based sampling paradigm in Hau et al. (2008). The purpose of the natural-mean heuristic is to calculate the mean of experienced observations for each prospect, then to choose the prospect with a larger experienced average outcome. Note that this model does not include the evaluation of value and probability weighting function as in the CPT-based models described in the previous section. The natural-mean heuristic is mathematically equivalent to the expected utility theory, though Hau et al. (2008) argued that the natural-mean heuristic may represent a different decision-making mechanism. The natural-mean heuristic works by summing all observations and dividing by the
number of observations. In other words, information about the experienced prospects is not processed in terms of probabilities, which makes the natural-mean heuristic different from the expected utility theory.

One problem of the natural-mean heuristic (and other heuristics) is that the heuristic would always choose a prospect with a larger natural mean. Consequently, there is no free parameter to account for the variability in the experimental data. This deterministic nature is incompatible with the hierarchical Bayesian framework. One solution includes adding a choice mechanism to the heuristic to solve the problem and create the stochastic structure to account for the random variability in data.

In reinforcement learning, two popular means to model binary or multinomial choices are ε-greedy choice rules and softmax choice rules (Sutton & Barto, 1998). The ε-greedy rules select the choice with the highest estimated payoff most of the time, but select an action at random with a small probability ε, independently of the estimated payoffs. The softmax rules, on the other hand, vary the choice probabilities as a graded function of estimated payoffs. The greedy option is still assigned the highest choice probability, but the others also are ranked and weighted according to the estimated payoffs. The two rules are essentially implementations of two choice mechanisms used in stochastic risky choice models, namely, the trembles model and the strong utility implemented as exponential Luce’s rule (Wilcox, 2008).

Figure 11 shows the graphical models for implementing the natural-mean heuristic with the choice mechanisms. The graphical model on the left panel applies the trembles stochastic model. In this model, individuals make a random choice with probability $\varepsilon_i$, and choose the prospect with the larger natural mean with probability $(1 -$
\( \varepsilon_i \). \( \varepsilon_i \) has its group-level hyper-distribution with hyper-parameters \( \mu_\varepsilon \) and \( \sigma_\varepsilon \). The probit transformation is used again, because \( \varepsilon_i \) is bounded between [0,1].

The natural means are calculated as \( NM_{ij} = p_{ij} \cdot r_{h,ij} + (1 - p_{ij}) \cdot r_{l,ij} \) for the risky prospect, and \( r_{m,ij} \) for the safe prospect. The latent probability \( \theta_{ij} \) of endorsing the risky prospect and the choice \( D_{ij} \) of the \( i^{th} \) subject on the \( j^{th} \) gamble game is defined as follows:

\[
\theta_{ij} = \begin{cases} 
1 - \frac{\varepsilon_i}{2} & \text{if } NM_{ij} > r_{m,ij} \\
\frac{\varepsilon_i}{2} & \text{if } NM_{ij} \leq r_{m,ij}
\end{cases} \quad \text{(link function)}
\]
The graphical model on the right panel applies the exponential Luce’s model. In this model, the latent probability $\theta_{ij}$ of endorsing the risky prospect is determined from the natural means of the prospects instead of the subjective values used in the CPT model. The other difference is that the sensitivity to the natural-mean difference $\phi_i$ is now treated as an individual parameter to account for the random variability in experimental data. $\phi_i$ has its group-level hyper-distribution with hyper-parameters $\mu_\phi$ and $\sigma_\phi$. The choice $D_{ij}$ is defined similarly as before:

$$\theta_{ij} = \frac{1}{1 + \exp(\phi_i(r_{m,ij} - NM_{ij}))} \quad \text{(link function)}$$

$$D_{ij} \sim \text{Bernoulli}(\theta_{ij}) \quad \text{(likelihood)}$$

There are relatively few studies in decision-making that compare between different choice mechanisms. Daw, O’Doherty, Dayan, Seymour, and Dolan (2006) compared the two methods in a four-armed bandit problem and suggested that the softmax methods (i.e., the exponential Luce’s rule) fitted the data better than the $\epsilon$-greedy methods (i.e., the trembles model). Under the description-based diagram, Stott (2006) compared the Luce’s rule, the exponential Luce’s rule, a constant error model and a probit model, concluding that the Luce’s rule and the exponential Luce’s rule performed better in terms of AIC. In the secondary data analysis, the exponential Luce’s rule is considered in both Studies 3 and 4, and the trembles model is considered in Study 4.
5.3 Study 3: HGM comparison of CPT-based models

Study 3 directly extends the simulation study in Study 2 by applying the same models to the data sets selected from the TPT (Erev et al., 2010). The purpose of Study 3 is to demonstrate the use of the hierarchical generalization modeling framework with empirical data. In addition, the analysis will provide some initial investigation on how the hierarchical generalization modeling framework can be used to test whether or not a decision-making process can generalize across different experimental paradigms.

5.3.1 Data summary

Two subsets of data from the first part of the TPT data (i.e., the estimation data set) were used in Study 3. The working data sets were obtained in two paradigms, decision-from-description (DFD) and decision-from-experience sampling paradigm (DFE-S). Only the gamble games with positive outcomes in the gaining domain were used, which consisted of one-third of the estimation data set in each paradigm. In the DFD data set, 20 subjects each completed 20 gamble games numbered 41-60 in the original data set. In the DFE-S data set, 40 subjects each completed 10 gamble games, half completed gambles 41-50 and half completed gambles 51-60. Table 14 summarizes the gamble designs and the proportion of choosing the risky prospect in the two data sets. Note that with in the gaining domain, the probabilities of a higher outcome are balanced among frequent events with higher probabilities ($p \geq .90$), rare events with lower probabilities ($p \leq .10$), and regular events with probabilities in-between.
Table 14. Gamble designs and the proportion of choosing the risky prospect in DFD and DFE-S paradigms.

<table>
<thead>
<tr>
<th>TPT #</th>
<th>( p_h )</th>
<th>( r_h )</th>
<th>( r_l )</th>
<th>( r_m )</th>
<th>DFD</th>
<th>DFE-S²</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>0.95</td>
<td>20.1</td>
<td>6.5</td>
<td>19.6</td>
<td>0.35</td>
<td>0.60</td>
</tr>
<tr>
<td>42</td>
<td>0.50</td>
<td>5.2</td>
<td>1.4</td>
<td>5.1</td>
<td>0.15</td>
<td>0.80</td>
</tr>
<tr>
<td>43</td>
<td>0.50</td>
<td>12.0</td>
<td>2.4</td>
<td>9.0</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td>44</td>
<td>0.90</td>
<td>20.7</td>
<td>9.1</td>
<td>19.8</td>
<td>0.20</td>
<td>0.70</td>
</tr>
<tr>
<td>45</td>
<td>0.07</td>
<td>8.4</td>
<td>1.2</td>
<td>1.6</td>
<td>0.55</td>
<td>0.70</td>
</tr>
<tr>
<td>46</td>
<td>0.40</td>
<td>22.6</td>
<td>7.2</td>
<td>12.4</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>47</td>
<td>0.93</td>
<td>23.4</td>
<td>7.6</td>
<td>22.1</td>
<td>0.65</td>
<td>0.20</td>
</tr>
<tr>
<td>48</td>
<td>0.09</td>
<td>17.2</td>
<td>5.0</td>
<td>5.9</td>
<td>0.10</td>
<td>0.75</td>
</tr>
<tr>
<td>49</td>
<td>0.90</td>
<td>18.9</td>
<td>6.7</td>
<td>17.7</td>
<td>0.05</td>
<td>0.45</td>
</tr>
<tr>
<td>50</td>
<td>0.04</td>
<td>12.8</td>
<td>4.7</td>
<td>4.9</td>
<td>0.70</td>
<td>0.20</td>
</tr>
<tr>
<td>51</td>
<td>0.03</td>
<td>19.1</td>
<td>4.8</td>
<td>5.2</td>
<td>0.15</td>
<td>0.65</td>
</tr>
<tr>
<td>52</td>
<td>0.91</td>
<td>12.3</td>
<td>1.3</td>
<td>12.1</td>
<td>0.35</td>
<td>0.25</td>
</tr>
<tr>
<td>53</td>
<td>0.90</td>
<td>6.8</td>
<td>3.0</td>
<td>6.7</td>
<td>0.40</td>
<td>0.25</td>
</tr>
<tr>
<td>54</td>
<td>0.30</td>
<td>22.6</td>
<td>9.2</td>
<td>11.0</td>
<td>0.85</td>
<td>0.25</td>
</tr>
<tr>
<td>55</td>
<td>0.09</td>
<td>6.4</td>
<td>0.5</td>
<td>1.5</td>
<td>0.15</td>
<td>0.70</td>
</tr>
<tr>
<td>56</td>
<td>0.06</td>
<td>15.3</td>
<td>5.9</td>
<td>7.1</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>57</td>
<td>0.90</td>
<td>5.3</td>
<td>1.5</td>
<td>4.7</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>58</td>
<td>0.50</td>
<td>21.9</td>
<td>8.1</td>
<td>12.6</td>
<td>0.25</td>
<td>0.65</td>
</tr>
<tr>
<td>59</td>
<td>0.70</td>
<td>27.5</td>
<td>9.2</td>
<td>21.9</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>60</td>
<td>0.20</td>
<td>4.4</td>
<td>0.7</td>
<td>1.1</td>
<td>0.20</td>
<td>0.55</td>
</tr>
</tbody>
</table>

¹ In DFD the probabilities were revealed to subjects. In DFE-S, the probabilities were used to determine the outcome show when the risky prospects were sampled. The experienced proportion of higher outcome \( r_h \) differed for each subject.

² The first 10 and last 10 gambles were complete by two different groups (N = 20 each).

5.3.2 Models

The competing models considered for the selected data sets are listed in Table 15. The value functions were fixed to be power value functions with individual parameter \( \alpha \). The probability weighting functions also were held constant for DFD data, and the TK weighting function was used with individual parameter \( \gamma \). For DFE-S, three candidate probability weighting functions were considered, including the TK weighting function.
(with parameter $\gamma$), the power weighting function (with parameter $\gamma$), and the linear weighting function (with parameter $w$). In terms of functional forms, the model in the first row of Table 15 is a fully generalizable model, whereas the models in the bottom rows are partially generalizable models.

The theoretical implications of these competing models are closely related with the cognitive processes underlying the decision-making mechanism in the decision-from-description and decision-from-experience sampling paradigms. The models are also associated with the investigation of what is known as the description-experience gap. The fully generalizable models assume no description-experience gap, or at least not in the functional forms of the value and probability weighting processes. The partially generalizable models, on the other hand, assume that the description-experience gap exists in terms of the different functional forms for the probability weighting process. Unlike previous approaches in the literature, the partially generalizable models also explicitly assume that some process, such as the value function, is generalizable across both decision-from-description and decision-from-experience sampling paradigms.

The other characteristics considered in setting up HGM is the invariance of parameter values. Two variations are considered: 1) invariant hyper-parameters for power value function (referred to as partial parameter invariance or PPI), i.e., $\mu_{a1} = \mu_{a2}$ and $\sigma_{a1} = \sigma_{a2}$; and 2) variant hyper-parameters for power value function (referred to as parameter variance or PV), i.e., $\mu_{a1} \neq \mu_{a2}$ and $\sigma_{a1} \neq \sigma_{a2}$. These two distinctions are interpreted in addition to the fully or partially generalizable hypotheses on functional forms. For example, the PPI version of the TK-linear model assumes not only the same functional
forms of value functions in DFD and DFE sampling paradigms, but also the same group-level parameters determining the parameter in evaluating the outcomes.

Table 15. Competing hierarchical generalization models for the selected TPT data sets.

<table>
<thead>
<tr>
<th>Models</th>
<th>Task1/DFD</th>
<th>Task 2/DFE-S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>value</td>
<td>weighting</td>
</tr>
<tr>
<td>TK-TK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TK-linear</td>
<td>power ($\alpha_1$)</td>
<td>TK ($\gamma_1$)</td>
</tr>
<tr>
<td>TK-power</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3.3 Results and discussion

Models are fit to data in JAGS by running two parallel chains for 5,000,000 iterations, with discarding the first 5,000 samples and setting the thinning interval at 50. The hyper-prior distributions are the same as in Study 2. Convergence is supported by various diagnostic measures. Table 16 compares the competing models with variant parameter values and lists the estimates of $\bar{D}$, $p_D$, and DIC values. Similarly, Table 17 compares the competing models with partially invariant parameter values. The smallest DIC value in each model set is underlined.

Model comparisons within each table suggest that the partially generalizable models outperform the fully generalizable model for the considered data set. The two partially generalizable models show a DIC difference of only around 1. Recall that in the model recovery simulation in Study 2 (Section 4.2.3), it was suggested that a reliable cutoff value for the DIC differences would be as low as around 10. Otherwise, it is reasonable to penalize less generalizable models and choose models with a larger degree of generalization. To investigate the potential sampling error in the estimation of DIC
values, a bootstrapping process is conducted in Study 4, and the discussion of a meaningful DIC difference is revisited then.

Table 16. Estimates of DIC related values for the competing model set with variant parameter values (PV).

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>824</td>
<td>36</td>
<td>860</td>
</tr>
<tr>
<td>TK-linear</td>
<td>809</td>
<td>46</td>
<td>855</td>
</tr>
<tr>
<td>TK-power</td>
<td>824</td>
<td>32</td>
<td>856</td>
</tr>
</tbody>
</table>

Table 17. Estimates of DIC related values for the competing model set with partially invariant parameter values (PPI).

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK-TK</td>
<td>820</td>
<td>33</td>
<td>853</td>
</tr>
<tr>
<td>TK-linear</td>
<td>805</td>
<td>43</td>
<td>848</td>
</tr>
<tr>
<td>TK-power</td>
<td>821</td>
<td>29</td>
<td>850</td>
</tr>
</tbody>
</table>

Before moving on to further analysis, it is worth comparing the PPI and PV forms of these models. This comparison is accomplished by comparing the DIC values from the second and the third rows in both tables, respectively. The DIC differences between the PPI and PV forms of each model are larger than the DIC differences between each model. Moreover, the PPI form assumes a more generalizable model than the PV form does. These two aspects seem to suggest that the models with partial invariant parameter values outperform the models with variant parameter values. Specifically, it suggests that the description-based data and the experience-based sampling data is better explained when assuming that the two paradigms share the same group level hyper-distributions for the power value function. In other words, it is reasonable to assume that the two paradigms at least share some of the underlying decision-making processes.
If using ranking of the DIC value as the exclusive model selection criterion, the best-fitting model will be the PPI form of TK-linear. The best-fitting model assumes a power value function for both decision-from-description and decision-from-experience sampling paradigms, TK probability weighting function for the description-based task, and linear weighting function for experience-based sampling task. Moreover, the power value functions share the same hyper-parameter distributions. Estimated posterior densities for the means of individual parameters are plotted in Figure 12. The estimated posterior densities for hyper-parameters \( \mu_{\alpha}, \mu_{\gamma_1}, \) and \( \mu_{w_2} \) are plotted. Only the means for the hyper-distributions are discussed here because these hyper-parameters are of more interest in interpretations. The black dotted lines indicate the means of the posterior samples. The gray dotted lines indicate the positions of 95% credible intervals. It is noteworthy that even with a small sample size, the width of estimated densities are comparable to those obtained in simulation studies and in previous studies (Nilsson et al., 2011), suggesting an acceptable model fit to the selected TPT data sets.

Generally speaking, the results from the above hierarchical generalization modeling comparison and analysis show that the procedure in Study 2 can be applied to the empirical data set in decision-from-description and decision-from-experience sampling paradigms. Using the hierarchical generalization modeling framework to analyze the TPT data offers a comprehensive investigation into the behavioral patterns shown in the description- and experience-based paradigms. Previous studies and the models in the TPT competition often used aggregated data to estimate the model parameters, and often assumed the same parameter values for a group of subjects in the same paradigm. The individual variability and the task variability can be separated in the
hierarchical generalization models. Additionally, it is possible to discover which process is more responsible for the variability. Given the PPI form of TK-linear is accepted as the “best” model in terms for the selected TPT data sets; it first showed the description-experience gap after accounting for the individual differences. Then it suggested that the differences in the two paradigms mainly derived from the process of probability weighting. Conversely, the processes of value evaluating are the same in the two paradigms and even share the same hyper-distributions. Therefore, it would be unnecessary to assume completely different mechanisms for the description- and experience-based paradigms.

Figure 12. Posterior samples of hyper-parameters of the model TK-Linear with partially invariant parameter values (PPI), which has the smallest DIC value in Study 3. (Top panel: $\mu_\alpha$; middle panel: $\mu_\gamma$; bottom panel: $\mu_\theta$. Black dotted lines: Posterior means of the hyper-parameters; gray dotted lines: 95% credible intervals.)
Note that the model comparison in Study 3 is designed within the CPT framework. However, previous studies of the decision-from-experience paradigms and the TPT competition show that models outside the CPT framework have better model-fitting and predictive performances. Checking the CPT-based models alone is not enough before reaching a reasonable conclusion. Study 4 continues with the hierarchical generalization modeling comparison by taking the heuristics into consideration.

**5.4 Study 4: HGM comparison between CPT and heuristics**

Study 4 extends the hierarchical generalization models examined in Study 3 to include more models that are discussed in the decision-from-experience literature. Instead of demonstrating the use of hierarchical generalization models, the focus of Study 4 shifts toward exploring the connection between model hypotheses and theoretical implications. The same data sets as in Study 3 are used. The summary of gamble designs and risk choice proportions are listed in Table 14.

**5.4.1 Models**

The hierarchical generalization models of foremost interest are still fully generalizable and partially generalizable models. In addition to the CPT-based models considered in Study 3, the natural-mean heuristic is included in Study 4. Therefore, an important change to the competing models is the addition of choice mechanisms as introduced in Section 5.2.2. To make the CPT-based models comparable with the natural-mean heuristic, the sensitivity parameter \( \varphi \) in the exponential Luce’s rule is no longer set as a tuning parameter. Instead, the sensitivity parameter \( \varphi \) is treated as a free parameter
following a hyper-distribution. The six competing hierarchical generalization models are listed in Table 18.

Table 18. Competing hierarchical generalization models for the selected TPT data sets.

<table>
<thead>
<tr>
<th>Models</th>
<th>Task1/DFD</th>
<th>Task 2/DFE-S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>utility</td>
<td>choice</td>
</tr>
<tr>
<td>CPT_Luce-CPT_Luce</td>
<td>CPT ($\gamma_1$)</td>
<td>Exponential Luce ($\phi_1$)</td>
</tr>
<tr>
<td>CPT_Luce-NM_Luce</td>
<td></td>
<td>CPT ($\gamma_2$)</td>
</tr>
<tr>
<td>CPT_Luce-CPT_tr</td>
<td></td>
<td>CPT ($\gamma_2$)</td>
</tr>
<tr>
<td>CPT_Luce-NM_tr</td>
<td></td>
<td>CPT ($\gamma_2$)</td>
</tr>
<tr>
<td>CPT_tr-CPT_tr</td>
<td>Trembles ($\epsilon_1$)</td>
<td>CPT ($\gamma_2$)</td>
</tr>
<tr>
<td>CPT_tr-NM_tr</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the initial analysis, treating the sensitivity parameter $\phi$ in the exponential Luce’s rule as a free, individual level parameter causes problems in fitting the CPT-based models. One potential problem, as noted in Study 3, is that the sample size is relatively small and cannot provide accurate estimates when the number of free parameters is large. Checking the tentative model fit suggests that the estimates for parameter $\alpha$ in the value function is close to 1. Considering that the scale for gamble outcomes is no larger than tens digit, and does not need approximation, Study 4 assumes a complete linear value function. That is, it is assumed that the subjects treat the outcome values as is, and the parameter $\alpha$ is not estimated in the CPT-based models.

Note that this ad hoc remedy should have little impact on the theoretical assumptions of interest. As reviewed in Chapter 2, the most discussed form of the description-experience gap is related to the probability weighting of rare events. No evidence has been documented, regarding the difference in evaluating the valuing of
gambles outcomes. On the other hand, when comparing CPT-based models and heuristics, the utility of a prospect is treated as a large unit of the model’s component. Thus, adjusting the value function within the CPT component does not alter the nature of the theoretical question under investigation.

5.4.2 Results and discussion

Table 19 summarize the DIC-related values from fitting the candidate models. An issue with the reported results in Table 19 is that some models have the problem of infinity when estimating the pD values. As presented in Table 19, the pD values for three problematic models are estimated using the medians instead of means. If the median estimates are acceptable, the models of CPT_Luce – CPT_tr (Row 3) and CPT_tr – CPT_tr (Row 5) have relatively smaller DIC values. Given that the infinite estimates for pD often suggest that the model fit is problematic, however, the CPT_tr – NM_tr model (Row 6) may be more appropriate for the data set. Another difficulty in interpreting the results in Table 19 is that the DIC values for either model pair differ by only 1. Thus it is hard to determine which one fits the data set better. In other words, it is unclear whether the DFD data is better explained by the CPT model with Luce’s choice rule, or by the CPT model with the trembles model.

To investigate the potential sampling error in the estimation of DIC values, a bootstrapping process (Efron, 1979) is conducted, as described in Table 20. A bootstrap sample size of 1,000 or larger usually is desired when estimating the distribution of a certain estimator. However, because fitting each model in Table 18 takes approximately 2 hours, a small size of 100 bootstrap samples is used, and only the CPT_Luce – NM_tr model (Row 4) and the CPT_tr – NM_tr model (Row 6) are fitted in the bootstrapping
process. The infinity problem in estimating $p_D$ occurs to many bootstrap samples in fitting both models. Hence, the median values of $p_D$ estimates are used to report the bootstrap DIC values. Results reveal that the DIC values for the CPT_Luce – NM_tr model range from 622 to 841, and the DIC values for the CPT_tr – NM_tr model range from 619 to 842. The bootstrap differences of DIC values, using the CPT_tr – NM_tr model as reference, are summarized in Table 21 and are plotted in Figure 13. The dotted line in Figure 13 indicates zero DIC difference, and the dashed line indicates the 95% bootstrap confidence interval. A negative value of the DIC difference indicates a preference of the CPT_tr – NM_tr model. Table 21 reveals that the mean DIC difference between the two models is -1.29, indicating a slight preference for the CPT_tr – NM_tr model. However, the relatively large standard error and the wide 95% bootstrap confidence interval suggests that the sampling error of the DIC values in the data is not negligible. With the size of the DIC difference in Table 19, it may be difficult to reach a conclusion.

Table 19. Estimates of the DIC related values for the competing model set.

<table>
<thead>
<tr>
<th>Models</th>
<th>$\bar{D}$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT_Luce – CPT_Luce</td>
<td>697</td>
<td>INF</td>
<td>N/A</td>
</tr>
<tr>
<td>CPT_Luce – NM_Luce</td>
<td>739</td>
<td>INF</td>
<td>N/A</td>
</tr>
<tr>
<td>CPT_Luce – CPT_tr</td>
<td>665</td>
<td>76*</td>
<td>741*</td>
</tr>
<tr>
<td>CPT_Luce – NM_tr</td>
<td>701</td>
<td>53*</td>
<td>753*</td>
</tr>
<tr>
<td>CPT_tr – CPT_tr</td>
<td>650</td>
<td>89*</td>
<td>739*</td>
</tr>
<tr>
<td>CPT_tr – NM_tr</td>
<td>686</td>
<td>67</td>
<td>753</td>
</tr>
</tbody>
</table>

* indicates that the $p_D$ value is estimated from median instead of mean because of the existence of infinite values. INF indicates that the estimate of $p_D$ has a problem of infinity.
Table 20. Pseudo code for bootstrapping process to estimate the DIC values.

1. For the description data set D.des, \(N_1 = 20\), assume \(P(n_1) = 1/20\);
   for the experience data set D.exp, \(N_2 = 39\), assume \(P(n_2) = 1/39\).
2. Sample \(n_1 = 20\) times from \(P(n_1)\), and form an index array \(\text{idx.des}[1:n1]\);
   sample \(n_2 = 39\) times from \(P(n_2)\), and form another index array \(\text{idx.exp}[1:n2]\).
3. Using the index arrays from Step 2, draw a bootstrap data sample from D.des and
   D.exp, respectively:
   
   ```
   for i in 1:n1 do  D.des.bs[i] = D.des[\text{idx.des}[i]]
   for j in 1:n2 do  D.exp.bs[j] = D.exp[\text{idx.exp}[j]]
   ```
4. Fit \{D.des.bs, D.exp.bs\} to models \{CPT_Luce – NM_tr, CPT_tr – NM_tr\}, and
   store the DIC related values.
5. Repeat Step 2 through Step 4 for \(M\) times, and get a bootstrapping distribution of
   the DIC values with size of \(M\).

Table 21. Summary statistics for the bootstrap DIC differences between the models of
CPT_Luce – NM_tr and CPT_tr – NM_tr.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Standard Error</th>
<th>Bootstrap CI 2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.29</td>
<td>-1.49</td>
<td>15.9</td>
<td>-24.9</td>
<td>29.4</td>
</tr>
</tbody>
</table>

Figure 13. Histogram of the bootstrap DIC differences between the CPT_Luce – NM_tr
and CPT_tr – NM_tr models.

(Dotted line: 0 difference in DIC values; dashed line: 95% bootstrap confidence interval.)
In terms of the generalization hypotheses, the CPT_tr – NM_tr model is a partially generalizable model that shares the functional form assumption of the choice mechanism. In contrast, the CPT_Luce – NM_tr model does not have any generalizable component. Therefore, although the two models have the same DIC values, the CPT_tr – NM_tr model may be preferred in favor of a generalizable component. Similarly, the CPT_Luce – CPT_tr model is a partially generalizable model, and the CPT_tr – CPT_tr model is a fully generalizable model that shares both functional form assumptions of probability weighting and choice mechanism. If the infinite pD problem is negligible, the CPT_tr – CPT_tr model would be preferred in spite of the small DIC differences.

A common characteristics of the preferable models is the assumption of the trembles model as the choice rule. The preference for the trembles model draws attention to a question that was not discussed much in studying the decision-making mechanism in the description- and experience-based paradigms. Previous studies tended to either treat choices as deterministic, or use Luce’s rules to retrieve the subjective values of the prospects and to adjust the latent probability (θ) that leads to the final choice. Both choice rules assume that the decision-makers are well attended during the experiment and are actively processing the information before making the choices. In contrast, the trembles model assumes that some proportion of the choices is made at random. For tedious experimental tasks or when the subjects are not highly motivated, it may be inevitable that the subjects occasionally make random choices. The trembles model can control for variability in the random responses, so that it does not confound with other decision-making processes in the model-fitting.
As many previous studies have already revealed, the CPT_tr – NM_tr model suggests that there is a major difference in the decision-making mechanisms between the description-based paradigm and the experience-based sampling paradigm. That is, when making choices in the description-based experiment, a subject tends to use the CPT model to evaluate the subjective value of the prospects, whereas, when making choices in the experience-based sampling experiment, a subject evaluates the prospects by the estimated natural means without weighting the probability information. On the other hand, a subject tends to make random choices without carefully evaluating the prospects for a certain proportion of choices. Hence, the description-experience gap may be generated from the different mechanisms used to evaluate the prospects, and it may be true that probability information is not important in the experience-based paradigm, as the results from the TPT competition suggest (Erev et al., 2010). Nevertheless, the two paradigms are not completely different. Subjects tend to use the same choice rule when determining the final choices in the description- and experience-based paradigms.

To further examine the model fits, the posterior estimates of the hyper-parameter distributions for the models of CPT_tr – CPT_tr and CPT_tr – NM_tr are plotted in Figure 14 and Figure 15, respectively. The black dotted lines indicate the means of the posterior samples. The gray dotted lines indicate the positions of 95% credible intervals. The two models have the same structural assumptions for the DFD task, and the corresponding hyper-parameters $\mu_{\gamma_1}$, $\sigma_{\gamma_1}$, $\mu_{\varepsilon_1}$, and $\sigma_{\varepsilon_1}$ display similar posterior distributions. On the other hand, the CPT_tr – CPT_tr model assumes the TK probability weighting function for the DFE-S task, whereas the CPT_tr – NM_tr model assumes the
natural-mean heuristic for the DFE-S task. Figure 14 depicts the posterior density for $\mu_{\gamma 2}$ and $\sigma_{\gamma 2}^2$, which are absent from Figure 15. $\mu_{\gamma 2}$ shows a distribution highly skewed toward 1, which differs from the distribution of $\mu_{\gamma 1}$. That is, even if the same probability weighting functions are assumed for DFD and DFE-S tasks in the CPT_tr – CPT_tr model, the parameters of the two processes would differ significantly, suggesting the existence of the description-experience gap. When compared with the CPT_tr – NM_tr model, the different assumptions on the DFE-S task also affect the estimates for hyperparameters $\mu_{\varepsilon 2}$, and $\sigma_{\varepsilon 2}^2$. The CPT_tr – NM_tr model suggests a larger proportion of random responses with a smaller variability within the group.

To summarize the findings in Studies 3 and 4, the results taken together seem to suggest that the hierarchical generalization modeling comparison and analysis show a description-experience gap in the underlying decision-making processes. Specifically, the probability information was processed differently in the decision-from-description paradigm and the decision-from-experience sampling paradigm. This result is obtained after accounting for individual variability and design variability in the TPT data by using the hierarchical generalization modeling framework, and hence, could be perceived as a supplementary evidence to the previous studies. What is new to the knowledge of the description-experience gap is that the choice rule may be generalizable across the different paradigms. This finding suggests that there is a common cognitive process involved in the decision-making mechanism for risky choice.
Figure 14. Posterior samples of hyper-parameters of the model CPT_tr – CPT_tr.
(Black dotted lines: posterior means; gray dotted lines: 95% credible intervals.)

Figure 15. Posterior samples of hyper-parameters of the model CPT_tr – NM_tr.
(Black dotted lines: posterior means; gray dotted lines: 95% credible intervals.)
Chapter 6: General Discussion

6.1 Executive summary

In recent decades, cognitive modeling has become a valuable tool for theory development and evaluation in cognitive science, and has proved useful in representing mental processes and understanding human behaviors (Busemeyer & Diederich, 2010; Lee & Wagenmakers, in press). Given that the development of cognitive models is based on a conceptual theoretical framework, it is desirable to systematically test the models with different theoretical hypotheses. The present dissertation is interested in developing a widely applicable yet statistically justified method to evaluate the ability of a cognitive model to generalize across different experimental tasks and conditions.

The framework of hierarchical generalization models is proposed to account for data observed in different experimental settings but were generated from the same cognitive processes. The hierarchical generalization models utilize a multilevel modeling approach and integrate data from various sources under one unified conceptual framework. If a certain cognitive process is generalizable, it may be represented in a hierarchical generalization model as a model component shared across experimental settings. The hierarchical generalization models are implemented using the hierarchical Bayesian modeling approach. This implementation takes advantage of the hierarchical
Bayesian modeling framework by appropriately accounting for individual differences and task variability, and in appropriately modeling nonlinear processes.

Another focus of this dissertation is to establish a statistically justified method to test a hierarchical generalization model in favor of the generalization hypothesis. Model check and model selection are two main methods to evaluate hierarchical Bayesian models, both of which have been discussed widely in the literature (for a comprehensive review, see Gelman, Meng, & Stern, 1996; Vehtari & Ojanen, 2012). The present dissertation employs the model selection approach and demonstrates the performance of the Deviance Information Criterion (DIC) for the purpose of checking the generalization hypotheses in the hierarchical generalization models. Furthermore, the implementation and evaluation of hierarchical generalization models were demonstrated in the context of behavioral decision-making to develop cognitive decision-making models using the decision-from-description and decision-from-experience paradigms.

Two simulation studies and two secondary data analysis studies were conducted. The simulation study in Study 1 demonstrated one application of hierarchical Bayesian modeling for cognitive decision-making models. Specifically, the Cumulative Prospect Theory (CPT) models with various probability weighting functions were used to generate simulated risky choice data. The simulated data were then fit to the hierarchical Bayesian model in order to check if the model structure that generated the data could be accurately recovered. The posterior densities of hyper-parameters showed the appropriate recovery of model parameters through the hierarchical Bayesian modeling. Study 2 was conducted to examine the performance of DIC in comparing a set of candidate hierarchical generalization models and in discovering the model that generated the simulation data.
Model recovery checks for different groups of candidate hierarchical generalization models suggested that DIC could generally discriminate the candidate models with different structural and parameter hypotheses. Note that the use of DIC was not without problems, given the ongoing debate in the literature arguing for and against DIC (e.g., the discussions following Spiegelhalter et al., 2002). Study 2 continued to check some problematic model pairs with relatively small differences in DIC. Estimation of DIC differences suggests that the DIC-based model selection is reliable and robust in general.

To summarize, the simulation studies in Study 1 and Study 2 have demonstrated the use of hierarchical Bayesian modeling and the DIC-based model selection to set up the framework of hierarchical generalization models.

Following the simulation studies, the secondary data analysis studies of Study 3 and Study 4 applied the framework of hierarchical generalization models to the empirical data observed in decision-from-description and decision-from-experience experiments. The data used in the secondary analysis were part of a comprehensive data set including data observed from one description-based task and two experience-based tasks. The comprehensive data set was originally collected and used for the purpose of a modeling competition known as the Technion Prediction Tournament (TPT, Erev et al., 2010). Studies 3 and 4 have used the data from description-based task and experience-based sampling task. Study 3 was an immediate follow-up of simulation studies, in that it used the same set of candidate hierarchical generalization models as those examined in Study 2. That is, the hierarchical generalization models considered in Study 3 all were based on the CPT framework, with a focus on the comparison of probability weighting functions. The DIC-based model selection results show a preference for hierarchical generalization
models with a generalizable modeling hypothesis for the parameter values. In terms of probability weighting functions, different structural assumptions better captured the patterns of data observed under different tasks. Nevertheless, the hierarchical generalization model that assumed the same functional form and the same group-level parameters outperformed the models that assumed different group-level parameters. The differences in the DIC values were not large enough to reach a conclusive decision. However, given the sample size and the extra design variability introduced by the procedure of experience-based sampling task, the results of Study 3 have at least shown a promising trend and would be informative in designing new experiments for comparing hierarchical generalization models.

Study 4 extended the spectrum of models considered in Study 3 to include decision-making models outside of the CPT framework. The new candidate model was a heuristic considered in previous model comparison studies of the decision-from-experience paradigms. The heuristic was combined with choice rules, to describe the stochastic process involved in the decision-making paradigms. The CPT-based models in Study 3 were still included, and also were combined with various choice rules in order to make the models comparable. Results from model-fitting and the DIC-based model selection again suggested different structural assumptions for the utility functions under description- and experience-based paradigms. With regard to the choice rules, there was some evidence favoring the generalizable assumption that the description-based paradigm and the experience-based sampling paradigm shared the same choice rule. Similar to the results from Study 3, the differences in the DIC values were not large enough to reach a conclusive decision, but showed a preference for a more generalizable hypothesis that the
description-based paradigm and the experience-based sampling paradigm share the same functional form and the same group level parameter values in the choice rule. The generalizable model hypotheses of the preferred hierarchical generalization model can be associated with the description-experience gap in risky decision-making. That is, based on the hierarchical generalization models, the TPT data showed a difference in evaluating the utilities of prospects, which was likely the main reason for the description-experience gap. However, there were still common cognitive processes in the different paradigms. The tasks may trigger the same choice mechanism even under different experimental settings. Consequently, more comprehensive models should be considered in developing theories of decision-making and predicting risky choice behaviors.

6.2 Main findings and implications

Based on the results and findings, one can draw several important implications for theoretical development and evaluation that highlight the main contributions of the present dissertation. Firstly, a hierarchical Bayesian modeling approach was adopted to develop a new statistical methodology for evaluating and comparing - among a set of cognitive models - in their ability to generalize across different experimental tasks and conditions. Secondly, the feasibility and soundness of the generalization methodology was demonstrated for testing various models of decision-making using data collected from decision-from-description and decision-from-experience experiments. In what follows, these contributions are discussed in detail.

6.2.1 Hierarchical Bayesian modeling for theory testing
The main methodological challenge and innovation of the present dissertation is to apply the hierarchical Bayesian modeling framework for the purpose of theory testing. Hierarchical Bayesian modeling was used in cognitive science to relate models of psychological processes to behavioral data (see Lee, 2011). Notwithstanding its potentials to account for various aspects of experimental data, the method has been used mainly to account for individual differences so far. The present dissertation has taken a step forward, proposing the framework of hierarchical generalization models to build models that express hypotheses of generalization. Within the framework of hierarchical generalization models, the notion of generalization was operationalized as the invariance of model structures and parameters.

Needless to say, there are many other uses of hierarchical Bayesian modeling in theory testing. For example, model development through iterative model building and model check has been used to build inductive inference models (for an extensive discussion on this issue, check Gelman & Shalizi, 2013 and the commentaries). Other than expressing hypotheses through functional forms and parameter values, informative priors also can be formalized to express theoretical hypotheses (Vanpaemel, 2011). These approaches are generally more appropriate in the early stage of theoretical exploration, whereas the model selection approach often performs better when there are already a few competing models for the cognitive process under investigation. Mixed modeling is a similar but alternative approach to model selection. A mixed model consists of several simpler models combined into a full hierarchical model, using a latent indicator parameter to choose between models and select one that is in effect for a certain condition (for an example in the hierarchical Bayesian modeling framework, see Shiffrin.
et al., 2008). Depending on the size of models and the nature of the theoretical question of interest, a different approach might work best. In general, however, the framework of hierarchical generalization models would be a good choice when data are already available from various experimental settings.

6.2.2 Performance of hierarchical generalization models

Having demonstrated the soundness of the hierarchical generalization models with simulated data, the methodology was applied in secondary data analyses with empirical data to address some important theoretical issues of a current debate in decision-making processes. The results show some moderate support for generalizable hypotheses, though not sufficiently convincing to draw firm conclusions. Aside from the aforementioned issues associated with DIC, the inconclusive results may be attributed to the method whereby the decision-from-experience experiments were conducted.

To elaborate, each subject experiences different information during the experiment. In some extreme cases, it may lead to a substantial change from the original gamble game. For example, if a certain risky prospect has a very small probability to yield a higher outcome, some subjects may not see the higher outcome during the whole sampling procedure. The corresponding gamble pair thus becomes a prospect with a lower outcome and a prospect with a medium outcome, and is no longer a risky choice situation for the subject. As a result, both task and experimental design differences are present in the data set. Although the hierarchical generalization models are able to model task, design, and individual variability separately, having too many sources of variability nonetheless have a negative impact on model-fitting. Furthermore, the TPT data set was not designed to be used for individual model-fitting. The number of subjects per problem
and the number of problems per task were smaller than the numbers used in simulation studies, which had a negative impact in the estimation of hyper-parameters. Although design differences should have been automatically considered in the framework of hierarchical generalization models, the DIC-based model selection was not optimized to account for such a difference. It seems that there are at least two issues with applying hierarchical generalization models in the secondary data analysis. Firstly, it is important to match the experimental designs with the methodological approach used for analyzing the data. Secondly, hierarchical generalization models should be improved to accommodate various practical problems before they can be widely used to unify data from different experimental settings.

6.3 Limitations and future work

As mentioned earlier, there are several limitations in the present dissertation, partly due to the constraints of time and computational load. These limitations point to some directions for future work to improve and refine the current framework of hierarchical generalization models.

First of all, the hierarchical generalization models were implemented in an open-source software package (JAGS), and the models are coded up using a predefined programming language for Bayesian graphical modeling (i.e., BUGS language). This particular choice of implementation method has its pros and cons. On one hand, using JAGS allows for the efficient implementation and evaluation of the method. It helped to speed up the computation and, at the same time, to avoid some common issues in MCMC simulation, such as using a block updating method for generalized linear models that
improves the mixing of Markov chains. The use of BUGS is popular in Bayesian communities. The BUGS-based implementation of the hierarchical generalization models would make it readily amenable to adaptations in cognitive science and psychology in general. On the other hand, relying on JAGS can limit the control over model fitting and model checking. For example, the use of DIC in the dissertation was partly a pragmatic choice since it is available in JAGS. This limitation can be overcome in the future even within JAGS by writing customized procedures.

Secondly, the observation that DIC is useful in the present dissertation does not necessarily suggest that DIC always should be used for the hierarchical generalization models. There is a growing body of work in the last two decades on the topic of Bayesian model selection. Accordingly, various model selection measures other than DIC have been proposed, such as the Bayes factor (Kass & Raftery, 1995; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008), the Widely Applicable Information Criterion (WAIC, Watanabe, 2010), and Bayesian cross-validation (for a review, see Vehtari & Ojanen, 2012). For hierarchical models in particular, there is no consensus on which measure is to be used under what condition, or whether model selection measures should be used at all. Some of the well discussed model selection methods, and model checking methods such as posterior predictive check, should be considered for use with the framework of hierarchical generalization models in future investigations.

Thirdly, note that the secondary data analysis studies are based on one particular set of experimental settings in one content area. Due to time constraints, only part of the TPT data was examined. Inclusion of gamble games with negative and mixed outcomes in the future would allow for a more comprehensive investigation into the model
behavior of the CPT and the description-experience gap. Furthermore, hierarchical generalization models were established with hierarchical Bayesian modeling framework in the dissertation, whereas the hierarchical Bayesian modeling framework was just recently introduced to the field of behavioral decision-making (Nilsson et al., 2011). Consequently, it is an open question whether or not other models for description- and experience-based tasks are suitable for the hierarchical Bayesian modeling framework, or how these models should be implemented. Future work in this direction can either extend the analysis to include more models, or modify hierarchical generalization models to allow for the use of non-Bayesian models. A final, long-term direction of future research will be to extend the scope of content areas to apply hierarchical generalization models, so as to explore the full potentials and merits of hierarchical generalization models as a way to assist theory testing.

6.4 Conclusion

In this dissertation, a statistical approach of hierarchical generalization modeling is developed, and the applicability of this approach is examined in the context of studies of behavioral decision-making. Results from simulation studies and secondary data analyses suggest that the hierarchical generalization modeling approach is theoretically justified and easily implementable in evaluating the generalizability of cognitive models. This approach supplements extant measures of generalizability, and expands the evaluation of generalizability to cover multiple experimental settings. In the long run, the hierarchical generalization modeling approach can be established as an effective methodology in testing cognitive theories.
References


Appendix A: Additional Simulation Results in Study 2
Table 22. DIC values estimated from fitting simulated data to the models with the TK probability weighting function \((N = 100, \sigma_{\mu_1} = \sigma_{\mu_2} = .01)\).

<table>
<thead>
<tr>
<th>Models</th>
<th>TK-TK</th>
<th>TK-Prelec I</th>
<th>TK-Power</th>
<th>TK-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from TK-TK</td>
<td>5135</td>
<td>5146</td>
<td>5191</td>
<td>5187</td>
</tr>
<tr>
<td>Data from TK-Prelec I</td>
<td>3949</td>
<td><strong>3793</strong></td>
<td>3827</td>
<td>3841</td>
</tr>
<tr>
<td>Data from TK-Power</td>
<td>5440</td>
<td>5353</td>
<td><strong>5241</strong></td>
<td>5277</td>
</tr>
<tr>
<td>Data from TK-Linear</td>
<td>5505</td>
<td>5427</td>
<td>4694</td>
<td><strong>4388</strong></td>
</tr>
</tbody>
</table>

* This is the same as Table 12 in Study 2, Section 4.2.2.

Table 23. DIC values estimated from fitting simulated data to the models with the TK probability weighting function \((N = 50, \sigma_{\mu_1} = \sigma_{\mu_2} = .01)\).

<table>
<thead>
<tr>
<th>Models</th>
<th>TK-TK</th>
<th>TK-Prelec I</th>
<th>TK-Power</th>
<th>TK-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from TK-TK</td>
<td>2562</td>
<td>2578</td>
<td>2603</td>
<td>2587</td>
</tr>
<tr>
<td>Data from TK-Prelec I</td>
<td>1937</td>
<td><strong>1856</strong></td>
<td>1863</td>
<td>1877</td>
</tr>
<tr>
<td>Data from TK-Power</td>
<td>2724</td>
<td>2662</td>
<td><strong>2603</strong></td>
<td>2612</td>
</tr>
<tr>
<td>Data from TK-Linear</td>
<td>2731</td>
<td>2704</td>
<td>2291</td>
<td><strong>2155</strong></td>
</tr>
</tbody>
</table>

Table 24. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function \((N = 100, \sigma_{\mu_1} = \sigma_{\mu_2} = .01)\).

<table>
<thead>
<tr>
<th>Models</th>
<th>Prl I-TK</th>
<th>Prl I-Prl I</th>
<th>Prl I-Power</th>
<th>Prl I-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Prl I-TK</td>
<td><strong>5208</strong></td>
<td>5222</td>
<td>5279</td>
<td>5265</td>
</tr>
<tr>
<td>Data from Prl I-Prl I</td>
<td>5368</td>
<td><strong>5366</strong></td>
<td>5388</td>
<td>5385</td>
</tr>
<tr>
<td>Data from Prl I-Power</td>
<td>5428</td>
<td>5476</td>
<td><strong>5312</strong></td>
<td>5338</td>
</tr>
<tr>
<td>Data from Prl I-Linear</td>
<td>5491</td>
<td>5541</td>
<td>4625</td>
<td><strong>4412</strong></td>
</tr>
</tbody>
</table>

Table 25. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function \((N = 50, \sigma_{\mu_1} = \sigma_{\mu_2} = .01)\).

<table>
<thead>
<tr>
<th>Models</th>
<th>Prl I-TK</th>
<th>Prl I-Prl I</th>
<th>Prl I-Power</th>
<th>Prl I-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Prl I-TK</td>
<td><strong>2598</strong></td>
<td>2609</td>
<td>2642</td>
<td>2635</td>
</tr>
<tr>
<td>Data from Prl I-Prl I</td>
<td>2691</td>
<td><strong>2685</strong></td>
<td>2703</td>
<td>2703</td>
</tr>
<tr>
<td>Data from Prl I-Power</td>
<td>2730</td>
<td>2732</td>
<td><strong>2678</strong></td>
<td>2705</td>
</tr>
<tr>
<td>Data from Prl I-Linear</td>
<td>2722</td>
<td>2763</td>
<td>2324</td>
<td><strong>2189</strong></td>
</tr>
</tbody>
</table>
Table 26. DIC values estimated from fitting simulated data to the models with the power probability weighting function ($N = 100, \sigma_{\mu\alpha} = \sigma_{\mu\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Power-TK</th>
<th>Power-Prl I</th>
<th>Power-Power</th>
<th>Power-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Power-TK</td>
<td><strong>5303</strong></td>
<td>5310</td>
<td>5434</td>
<td>5344</td>
</tr>
<tr>
<td>Data from Power-Prl I</td>
<td>5283</td>
<td><strong>5286</strong></td>
<td>5428</td>
<td>5318</td>
</tr>
<tr>
<td>Data from Power-Power</td>
<td>5474</td>
<td>5476</td>
<td><strong>5376</strong></td>
<td>5416</td>
</tr>
<tr>
<td>Data from Power-Linear</td>
<td>5480</td>
<td>5482</td>
<td>4762</td>
<td><strong>4357</strong></td>
</tr>
</tbody>
</table>

Table 27. DIC values estimated from fitting simulated data to the models with the power probability weighting function ($N = 50, \sigma_{\mu\alpha} = \sigma_{\mu\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Power-TK</th>
<th>Power-Prl I</th>
<th>Power-Power</th>
<th>Power-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Power-TK</td>
<td><strong>2626</strong></td>
<td>2630</td>
<td>2704</td>
<td>2659</td>
</tr>
<tr>
<td>Data from Power-Prl I</td>
<td>2705</td>
<td><strong>2713</strong></td>
<td>2750</td>
<td>2712</td>
</tr>
<tr>
<td>Data from Power-Power</td>
<td>2722</td>
<td>2722</td>
<td><strong>2660</strong></td>
<td>2682</td>
</tr>
<tr>
<td>Data from Power-Linear</td>
<td>2761</td>
<td>2762</td>
<td>INF</td>
<td><strong>2294</strong></td>
</tr>
</tbody>
</table>

Note: INF indicates that the estimate of $p_D$ has a problem of infinity.

Table 28. DIC values estimated from fitting simulated data to the models with the linear probability weighting function ($N = 100, \sigma_{\mu\alpha} = \sigma_{\mu\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Linear-TK</th>
<th>Linear-Prl I</th>
<th>Linear-Power</th>
<th>Linear-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Linear-TK</td>
<td><strong>4385</strong></td>
<td>4411</td>
<td>4459</td>
<td>4559</td>
</tr>
<tr>
<td>Data from Linear-Prl I</td>
<td>4379</td>
<td><strong>4374</strong></td>
<td>4405</td>
<td>4557</td>
</tr>
<tr>
<td>Data from Linear-Power</td>
<td>4590</td>
<td>4592</td>
<td><strong>4505</strong></td>
<td>4652</td>
</tr>
<tr>
<td>Data from Linear-Linear</td>
<td>4729</td>
<td>4729</td>
<td>3868</td>
<td><strong>3591</strong></td>
</tr>
</tbody>
</table>

Table 29. DIC values estimated from fitting simulated data to the models with the linear probability weighting function ($N = 50, \sigma_{\mu\alpha} = \sigma_{\mu\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Linear-TK</th>
<th>Linear-Prl I</th>
<th>Linear-Power</th>
<th>Linear-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Linear-TK</td>
<td><strong>2170</strong></td>
<td>2175</td>
<td>2196</td>
<td>INF</td>
</tr>
<tr>
<td>Data from Linear-Prl I</td>
<td>2213</td>
<td><strong>2205</strong></td>
<td>2240</td>
<td>2302</td>
</tr>
<tr>
<td>Data from Linear-Power</td>
<td>2256</td>
<td>2257</td>
<td><strong>2172</strong></td>
<td>INF</td>
</tr>
<tr>
<td>Data from Linear-Linear</td>
<td>2328</td>
<td>2329</td>
<td>1926</td>
<td><strong>1834</strong></td>
</tr>
</tbody>
</table>

Note: INF indicates that the estimate of $p_D$ has a problem of infinity.
Table 30. DIC values estimated from fitting simulated data to the models with the TK probability weighting function (N = 100, $\sigma_{\mu} = \sigma_{\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>TK-TK</th>
<th>TK-Prelec</th>
<th>TK-Power</th>
<th>TK-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from TK-TK</td>
<td>2567</td>
<td>2815</td>
<td>3312</td>
<td>3007</td>
</tr>
<tr>
<td>Data from TK-Prelec</td>
<td>1997</td>
<td>1578</td>
<td>2812</td>
<td>2668</td>
</tr>
<tr>
<td>Data from TK-Power</td>
<td>3005</td>
<td>3102</td>
<td>2562</td>
<td>2934</td>
</tr>
<tr>
<td>Data from TK-Linear</td>
<td>3077</td>
<td>3091</td>
<td>3101</td>
<td>2718</td>
</tr>
</tbody>
</table>

Note: Tables 30 to 33 are based on the following set of data-generating parameters. For the power value function, the data-generating hyper-parameter is $\mu_a = .88$. For the probability weighting functions, the data-generating hyper-parameters are: $\mu_\gamma = .53$ for the TK function; $\mu_\gamma = .61$ for the Prelec I function; $\mu_\gamma = 1.7$ for the power function; and $\mu_w = .9$ for the linear function.

Table 31. DIC values estimated from fitting simulated data to the models with the Prelec I probability weighting function (N = 100, $\sigma_{\mu} = \sigma_{\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Prl I-TK</th>
<th>Prl I-Prl I</th>
<th>Prl I-Power</th>
<th>Prl I-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Prl I-TK</td>
<td>2573</td>
<td>2829</td>
<td>3396</td>
<td>3062</td>
</tr>
<tr>
<td>Data from Prl I-Prl I</td>
<td>2997</td>
<td>2816</td>
<td>4026</td>
<td>3836</td>
</tr>
<tr>
<td>Data from Prl I-Power</td>
<td>3196</td>
<td>3583</td>
<td>2777</td>
<td>3184</td>
</tr>
<tr>
<td>Data from Prl I-Linear</td>
<td>3014</td>
<td>3226</td>
<td>3047</td>
<td>2627</td>
</tr>
</tbody>
</table>

Table 32. DIC values estimated from fitting simulated data to the models with the power probability weighting function (N = 100, $\sigma_{\mu} = \sigma_{\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Power-TK</th>
<th>Power-Prl I</th>
<th>Power-Power</th>
<th>Power-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Power-TK</td>
<td>2540</td>
<td>2650</td>
<td>3438</td>
<td>3064</td>
</tr>
<tr>
<td>Data from Power-Prl I</td>
<td>2934</td>
<td>2761</td>
<td>4030</td>
<td>3764</td>
</tr>
<tr>
<td>Data from Power-Power</td>
<td>3170</td>
<td>3411</td>
<td>2787</td>
<td>3183</td>
</tr>
<tr>
<td>Data from Power-Linear</td>
<td>2925</td>
<td>2971</td>
<td>2996</td>
<td>2638</td>
</tr>
</tbody>
</table>

Table 33. DIC values estimated from fitting simulated data to the models with the linear probability weighting function (N = 100, $\sigma_{\mu} = \sigma_{\gamma} = .01$).

<table>
<thead>
<tr>
<th>Models</th>
<th>Linear-TK</th>
<th>Linear-Prl I</th>
<th>Linear-Power</th>
<th>Linear-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data from Linear-TK</td>
<td>2521</td>
<td>2665</td>
<td>3319</td>
<td>3049</td>
</tr>
<tr>
<td>Data from Linear-Prl I</td>
<td>2865</td>
<td>2679</td>
<td>3917</td>
<td>3839</td>
</tr>
<tr>
<td>Data from Linear-Power</td>
<td>3083</td>
<td>3352</td>
<td>2729</td>
<td>3113</td>
</tr>
<tr>
<td>Data from Linear-Linear</td>
<td>3017</td>
<td>3084</td>
<td>3063</td>
<td>2683</td>
</tr>
</tbody>
</table>
Appendix B: Scripts for R and JAGS
B.1 Example of data generation

# Task 1: T-K probability weighting and power value function
# Task 2: T-K probability weighting and power value function
# Hyper-parameters={alpha, gamma}*{mu, sigma}
# Gamble design={prob.h, reward.h, reward.l, reward.m, Domain}

gen.1212 <- function(eps, hyper, N.subj, design=as.data.frame(prospects))
{
  design <- as.data.frame(design)
  attach(design)
  N.gam <- dim(design)[1]

  rh.mat <- array(rh, dim=c(N.gam, N.subj))
  rl.mat <- array(rl, dim=c(N.gam, N.subj))
  rm.mat <- array(rm, dim=c(N.gam, N.subj))
  ph.mat <- array(ph, dim=c(N.gam, N.subj))
  v.rh <- v.rl <- v.rm <- array(0, dim=c(N.gam, N.subj, 2))
  w.rh <- w.rl <- array(0, dim=c(N.gam, N.subj, 2))

  # Generating Individual Parameters
  ind.parm <- array(rnorm(2*N.subj*2, hyper[,1], hyper[,2]),
                   dim=c(4, N.subj))

  #--------- Task 1 ---------
  ind.a <- t(array(ind.parm[1,], c(N.subj, N.gam)))
  ind.g <- t(array(ind.parm[2,], c(N.subj, N.gam)))

  # Value function: Power
  v.rh[rh>=0,,1] <- rh.mat[rh>=0,]^ind.a[rh>=0]
  v.rl[rl>=0,,1] <- rl.mat[rl>=0,]^ind.a[rl>=0]
  v.rm[rm>=0,,1] <- rm.mat[rm>=0,]^ind.a[rm>=0]

  # Probability Weighting: TK
  w.r.z <- (ph.mat[,]^ind.g) + ((1-ph.mat[,])^ind.g) ^ (1/ind.g)
  w.rh[,,1] <- (ph.mat[,]^ind.g)/w.r.z
  w.rl[,,1] <- ((1-ph.mat[,])^ind.g)/w.r.z

  #--------- Task 2 ---------
  ind.a <- t(array(ind.parm[3,], c(N.subj, N.gam)))
  ind.g <- t(array(ind.parm[4,], c(N.subj, N.gam)))

  # Value function: Power
  v.rh[rh>=0,,2] <- rh.mat[rh>=0,]^ind.a[rh>=0]
  v.rl[rl>=0,,2] <- rl.mat[rl>=0,]^ind.a[rl>=0]
  v.rm[rm>=0,,2] <- rm.mat[rm>=0,]^ind.a[rm>=0]

  # Probability Weighting: TK
w.r.z <- ((ph.mat[,]^ind.g) + ((1-ph.mat[,]]^ind.g))^ (1/ind.g)
w.rh[,2] <- (ph.mat[,]^ind.g)/w.r.z
w.rl[,2] <- ((1-ph.mat[,]]^ind.g)/w.r.z

#-------- Bernoulli Decisions for Task 1 and 2 --------
# SV and theta from Luce's choice rule
sv.r <- w.rh * v.rh + w.rl * v.rl
sv.s <- v.rm
theta <- 1/(1+exp(eps * (sv.s-sv.r)))

# Bernoulli trials
dvec <- rbinom(N.gam*N.subj*2,1,theta)

# Reshape the generated choice data
# Dim 1 (row): gambles
# Dim 2 (column): subjects
d <- array(dvec, dim=c(N.gam,N.subj,2))

detach(design)
return(list(choice=d, theta=theta, ind=ind.parm, hyper=hyper))
}
B.2 Example of fitting simulation data and calculating DIC values

#-------- Parallelize --------
suppressMessages(library(foreach))
suppressMessages(library(parallel))
suppressMessages(library(doParallel))

#-------- Load rjags and modules --------
suppressMessages(library(rjags))
load.module('glm')
load.module('dic')
load.module('lecuyer')

#-------- Read Gamble Designs --------
wdir <- 'C:/ simulation/
setwd(wdir)
source('./source/ReadDesign.r')

#-------- Call Data-Generating Routine --------
# Hyper-parameters={alpha, gamma}*{mu,sigma}
# Gamble design={prob.h, reward.h, reward.l, reward.m, Domain}
# Parameters to serve: eps, hyper, N.subj,
# design=as.data.frame(prospects)
source('./source/DataGen-pow-TK.r')
N.gam <- dim(prospects)[1]
eps <- 3
hypvar <- .01
hyper.1 <- matrix(c(.48,.74,hypvar,hypvar),nrow=2)
hyper.2 <- matrix(c(.48,.71,hypvar,hypvar),nrow=2)
hyper.3 <- matrix(c(.48,.84,hypvar,hypvar),nrow=2)
hyper.4 <- matrix(c(.48,1.2,hypvar,hypvar),nrow=2)

#-------- Simulating Data --------
N.subj <- 100
d.1 <- gen.1211(eps, rbind(hyper.2,hyper.1), N.subj)
d.2 <- gen.1212(eps, rbind(hyper.2,hyper.2), N.subj)
d.3 <- gen.1213(eps, rbind(hyper.2,hyper.3), N.subj)
d.4 <- gen.1214(eps, rbind(hyper.2,hyper.4), N.subj)
data.sim <- list(d.1, d.2, d.3, d.4)

#-------- Model Settings --------
node.DIC <- c("deviance","pD")
para.1 <- c(node.DIC, "mu.alpha", "mu.gamma1", "mu.gamma2")
para.2 <- c(node.DIC, "mu.alpha", "mu.gamma")
para.3 <- c(node.DIC, "mu.alpha", "mu.gamma1", "mu.g2")
para.4 <- c(node.DIC, "mu.alpha", "mu.gamma1", "mu.wght2")
para.sim <- list(para.1, para.2, para.2pv, para.3, para.4)
# Model.N = Number of models to be fitted in each core
# Sim.N = Number of data-generating models
n.chain <- 2
Model.N <- length(model.sim)
Sim.N <- length(data.sim)
rngInits <- parallel.seeds("lecuyer::RngStream",
n.chain*Sim.N*Model.N)
n.Iter <- 1e5
n.BurnIn <- 5e3

cl <- makeCluster(Sim.N)
registerDoParallel(cl)
dir.create('./bkp/')
res <- foreach(i=1:Sim.N,.packages=c("rjags")) %dopar% {
  load.module('glm')
  load.module('dic')
  load.module('lecuyer')

  #--------------------- Prepare data ---------------------
data.jags <- list("d1.data"=data.sim[[i]]$choice[,,1],
    "d2.data"=data.sim[[i]]$choice[,,2], "prospects"=prospects,
    "N.subj"=c(N.subj, N.subj), "eps"=eps)
  res.grp <- vector("list", Model.N)
  Sim.DIC <- array(0, dim=c(Model.N)
  colnames(Sim.DIC) <- c("DIC","pD","deviance","var(dev)")

  #--------------------- Fitting HGM ---------------------
  for(j in 1:Model.N) {
    m.j <- jags.model(file=file.path(paste(wdir, '/model/','sep=""'), model.sim[[j]]), data=data.jags,
      n.chain=n.chain, n.adapt=100, inits=rngInits[(i-1)*n.chain*Model.N+(j-1)*n.chain+(1:n.chain)])
    update(m.j, n.BurnIn)
    res.j <- jags.samples(m.j, para.sim[[j]], n_iter)
    Sim.DIC[j,] <- c(mean(res.j$deviance)+mean(res.j$pD),
      mean(res.j$pD), mean(res.j$deviance),
      var(res.j$deviance))
    res.grp[[j]] <- res.j
  }
  return(list('Sim.DIC'=Sim.DIC, 'samples'=res.grp))
}
stopCluster(cl)
B.3 Examples of CPT-based hierarchical generalization models

*m-pow-TK-pow-TK.txt*

# Task 1: T-K probability weighting and power value function
# Task 2: T-K probability weighting and power value function
# Use positive rewards only
model {
    #------- Prior Specification -------
    # Priors for individual-level parameters, alpha and beta
    mapped on to [0, 1] using probit function
    for(j in 1:N.subj[1])
    {
        alpha1[j] <- phi(alpha1.phi[j])
        alpha1.phi[j] ~ dnorm(mu.phi.alpha,tau.phi.alpha)T(-3, 3)
        gamma1[j] <- phi(gamma1.phi[j])
        gamma1.phi[j] ~ dnorm(mu.phi.gamma1,tau.phi.gamma1)T(-3, 3)
    }
    for(j in 1:N.subj[2])
    {
        alpha2[j] <- phi(alpha2.phi[j])
        alpha2.phi[j] ~ dnorm(mu.phi.alpha,tau.phi.alpha)T(-3, 3)
        gamma2[j] <- phi(gamma2.phi[j])
        gamma2.phi[j] ~ dnorm(mu.phi.gamma2,tau.phi.gamma2)T(-3, 3)
    }
    # Priors for group-level hyper-parameters
    mu.phi.alpha ~ dnorm(0,1)
    tau.phi.alpha <- pow(sigma.phi.alpha,-2)
    sigma.phi.alpha ~ dunif(0,10)
    mu.alpha <- phi(mu.phi.alpha)
    mu.phi.gamma1 ~ dnorm(0,1)
    tau.phi.gamma1 <- pow(sigma.phi.gamma1,-2)
    sigma.phi.gamma1 ~ dunif(0,10)
    mu.gamma1 <- phi(mu.phi.gamma1)
    mu.phi.gamma2 ~ dnorm(0,1)
    tau.phi.gamma2 <- pow(sigma.phi.gamma2,-2)
    sigma.phi.gamma2 ~ dunif(0,10)
    mu.gamma2 <- phi(mu.phi.gamma2)
### Model Specification and Fitting

for(j in 1:N.subj[1]) # Subject-loop, Task 1
{
  for(i in 1:20) # Item-loop
  {
    ## Risky prospects
    # Value function: Power
    v1.rh[i,j] <- pow(prospects[i,2], alpha1[j])
    v1.rl[i,j] <- pow(prospects[i,3], alpha1[j])

    # Probability weighting function: TK
    w1.rh[i,j] <- pow(prospects[i,1], gamma1[j]) / w1.r.z[i,j]
    w1.rl[i,j] <- pow((1-prospects[i,1]), gamma1[j]) / w1.r.z[i,j]
    w1.r.z[i,j] <- pow(pow(prospects[i,1], gamma1[j]) + pow((1-prospects[i,1]), gamma1[j]), (1/gamma1[j]))

    ## Safe prospects
    # Value function: Power
    v1.rm[i,j] <- pow(prospects[i,4], alpha1[j])

    ## Bernoulli decisions
    # Subjective value
    sv1.r[i,j] <- w1.rh[i,j] * v1.rh[i,j] + w1.rl[i,j] * v1.rl[i,j]
    sv1.s[i,j] <- v1.rm[i,j]

    # theta from Luce's choice rule
    binval1[i,j] <- 1 / (1 + exp(eps * (sv1.s[i,j] - sv1.r[i,j])))
    d1.data[i,j] ~ dbern(binval1[i,j])
  } # End Item-loop
} # End Subject-loop, Task 1

for(j in 1:N.subj[2]) # Subject-loop, Task 2
{
  for(i in 1:20) # Item-loop
  {
    ## Risky prospects
    # Value function: Power
    v2.rh[i,j] <- pow(prospects[i,2], alpha2[j])
    v2.rl[i,j] <- pow(prospects[i,3], alpha2[j])

    # Probability weighting function: TK
    w2.rh[i,j] <- pow(prospects[i,1], gamma2[j]) / w2.r.z[i,j]
    w2.rl[i,j] <- pow((1-prospects[i,1]), gamma2[j]) / w2.r.z[i,j]

    # Subjective value
    sv2.r[i,j] <- w2.rh[i,j] * v2.rh[i,j] + w2.rl[i,j] * v2.rl[i,j]
    sv2.s[i,j] <- v2.rm[i,j]

    # theta from Luce's choice rule
    binval2[i,j] <- 1 / (1 + exp(eps * (sv2.s[i,j] - sv2.r[i,j])))
    d2.data[i,j] ~ dbern(binval2[i,j])
  } # End Item-loop
} # End Subject-loop, Task 2
w2.r.z[i,j] <- pow(pow(prospects[i,1],gamma2[j])
+ pow((1-prospects[i,1]),gamma2[j]),
(1/gamma2[j]))

## Safe prospects
# Value function: Power
v2.rm[i,j] <- pow(prospects[i,4], alpha2[j])

## Bernoulli decisions
# Subjective value
sv2.r[i,j] <- w2.rh[i,j] * v2.rh[i,j] +
w2.rl[i,j] * v2.rl[i,j]
sv2.s[i,j] <- v2.rm[i,j]

# theta from Luce's choice rule
binval2[i,j] <- 1 / (1 + exp(eps * (sv2.s[i,j] -
sv2.r[i,j])))
d2.data[i,j] ~ dbern(binval2[i,j])
}
} # End Item-loop
} # End Subject-loop, Task 2
} # End of model
# Task 1: T-K probability weighting and power value function
# Task 2: Power probability weighting and power value function
# Use positive rewards only

def model:
    #------- Prior Specification -------
    # Priors for individual-level parameters, alpha and beta
    mapped on to [0, 1] using probit function
    for (j in 1:N.subj[1])
    {
        alpha1[j] <- phi(alpha1.phi[j])
        alpha1.phi[j] ~ dnorm(mu.phi.alpha, tau.phi.alpha) T(-3, 3)
        gamma1[j] <- phi(gamma1.phi[j])
        gamma1.phi[j] ~ dnorm(mu.phi.gamma1, tau.phi.gamma1) T(-3, 3)
    }
    for (j in 1:N.subj[2])
    {
        alpha2[j] <- phi(alpha2.phi[j])
        alpha2.phi[j] ~ dnorm(mu.phi.alpha, tau.phi.alpha) T(-3, 3)
        g2[j] ~ dnorm(mu.g2, tau.g2) T(0.1, 10)
    }
    # Priors for group-level hyper-parameters
    mu.phi.alpha ~ dnorm(0, 1)
    tau.phi.alpha <- pow(sigma.phi.alpha, -2)
    sigma.phi.alpha ~ dunif(0, 10)
    mu.alpha <- phi(mu.phi.alpha)
    
    mu.phi.gamma1 ~ dnorm(0, 1)
    tau.phi.gamma1 <- pow(sigma.phi.gamma1, -2)
    sigma.phi.gamma1 ~ dunif(0, 10)
    mu.gamma1 <- phi(mu.phi.gamma1)
    
    mu.g2 ~ dnorm(0, 1)
    tau.g2 <- pow(sigma.g2, -2)
    sigma.g2 ~ dunif(0, 10)
# Model Specification and Fitting

for(j in 1:Nsubj[1]) # Subject-loop, Task 1
{
    for(i in 1:20) # Item-loop
    {
        ## Risky prospects
        # Value function: Power
        v1.rh[i,j] <- pow(prospects[i,2], alpha1[j])
        v1.rl[i,j] <- pow(prospects[i,3], alpha1[j])

        # Probability weighting function: TK
        w1.rh[i,j] <- pow(prospects[i,1], gamma1[j]) / w1.r.z[i,j]
        w1.rl[i,j] <- pow((1-prospects[i,1]), gamma1[j]) / w1.r.z[i,j]
        w1.r.z[i,j] <- pow(pow(prospects[i,1], gamma1[j]) + pow((1-prospects[i,1]), gamma1[j]), (1/gamma1[j]))

        ## Safe prospects
        # Value function: Power
        v1.rm[i,j] <- pow(prospects[i,4], alpha1[j])

        ## Bernoulli decisions
        # Subjective value
        sv1.r[i,j] <- w1.rh[i,j] * v1.rh[i,j] + w1.rl[i,j] * v1.rl[i,j]
        sv1.s[i,j] <- v1.rm[i,j]

        # theta from Luce's choice rule
        binval1[i,j] <- 1 / (1 + exp(eps * (sv1.s[i,j] - sv1.r[i,j])))
        d1.data[i,j] ~ dbern(binval1[i,j])
    } # End Item-loop
} # End Subject-loop, Task 1

for(j in 1:Nsubj[2]) # Subject-loop, Task 2
{
    for(i in 1:20) # Item-loop
    {
        ## Risky prospects
        # Value function: Power
        v2.rh[i,j] <- pow(prospects[i,2], alpha2[j])
        v2.rl[i,j] <- pow(prospects[i,3], alpha2[j])

        # Probability weighting function: Power
        w2.rh[i,j] <- pow(prospects[i,1], g2[j])
        w2.rl[i,j] <- pow((1-prospects[i,1]), g2[j])

        ## Safe prospects
# Value function: Power
v2.rm[i,j] <- pow(prospects[i,4], alpha2[j])

## Bernoulli decisions
# Subjective value
sv2.r[i,j] <- w2.rh[i,j] * v2.rh[i,j] + w2.rl[i,j] * v2.rl[i,j]
sv2.s[i,j] <- v2.rm[i,j]

# theta from Luce's choice rule
binval2[i,j] <- 1 / (1 + exp(eps * (sv2.s[i,j] - sv2.r[i,j])))
d2.data[i,j] ~ dbern(binval2[i,j])

} # End Item-loop

} # End Subject-loop, Taks 2

} # End of model
B.4 Example of fitting TPT data and calculating DIC values

#-------- Load rjags and modules --------
suppressMessages(library(rjags))
load.module('glm')
load.module('dic')
load.module('lecuyer')

#-------- Prepare Data --------
wdir <- 'C:/simulation/
setwd(wdir)
source('./source/LoadTPT-pos.r')

# For CPT models
data.tk <- tpt.pos
data.tk$ev1.r <- data.tk$ev1.s <- NULL
data.tk$ev2.r <- data.tk$ev2.s <- NULL
data.tk1 <- tpt.pos[c("N.subj","N.gamma","d1.data","prospects.1")]data.tk2 <- tpt.pos[c("d2.data", "prospects.2")]  

# For Natural-Mean with trembles model
data.nmt <- tpt.pos[c("d1.data", "d2.data", "N.subj", "N.gamma")]data.nmt$dnm.1 <- (tpt.pos$ev1.r > tpt.pos$ev1.s)mode(data.nmt$dnm.1) <- "numeric"data.nmt$dnm.2 <- (tpt.pos$ev2.r > tpt.pos$ev2.s)mode(data.nmt$dnm.2) <- "numeric"
data.nmt2 <- tpt.pos["d2.data"]data.nmt2$dnm.2 <- (tpt.pos$ev2.r > tpt.pos$ev2.s)mode(data.nmt2$dnm.2) <- "numeric"

# For Natural-Mean with exponential Luce model
data.nml <- tpt.pos
data.nml$prospects.1 <- data.nml$prospects.2 <- NULL
data.nml2 <- tpt.pos[c("d2.data", "ev2.r", "ev2.s")]  

#-------- Pulling up Models --------
node.DIC <- c("deviance","pD")
para <- list("para.tkl"=c(node.DIC, "mu.gamma1", "mu.gamma2", 
"mu.luce1", "mu.luce2"),
"para.tkl.tktr"=c(node.DIC, "mu.gamma1", 
"mu.gamma2", "mu.luce1", "mu.eps2"),
"para.tkl.nml"=c(node.DIC, "mu.gamma1", "mu.luce1", 
"mu.luce2"),
"para.tkl.nmtr"=c(node.DIC, "mu.gamma1", "mu.luce1", 
"mu.eps2"),
"para.tktr.tktr"=c(node.DIC, "mu.gamma1", "mu.gamma2", 
"mu.luce1", "mu.eps1", "mu.eps2"),
"para.tktr.nmtr"=c(node.DIC, "mu.gamma1", "mu.eps1", "mu.eps2"))
model <- list("CPT_luce-CPT_luce.txt", "CPT_luce-CPT_tr.txt", "CPT_luce-nm_luce.txt", "CPT_luce-nm_tr.txt", "CPT_tr-CPT_tr.txt", "CPT_tr-nm_tr.txt")
data.jags <- list("data.tkl"=data.tk, "data.tkl.tktr"=data.tk, "data.tkl.nml"=c(data.tk1, data.nml2), "data.tktr.nmtr"=c(data.tk1, data.nmt2), "data.tktr.tktr"=data.tk, "data.tktr.nmtr"=c(data.tk1, data.nmt2))

#------------------------ Fitting HGM and Calculating DIC ------------------------
n.chain <- 2
n.Iter <- 1e6
n.BurnIn <- 5e3
n.Thin <- 50
Model.N <- length(model)
rngInits <- parallel.seeds("lecuyer::RngStream", n.chain*Model.N)
Sim.DIC <- array(0, dim=c(Model.N, 5))
colnames(Sim.DIC) <- c("DIC", "pD(Md)", "pD(Mn)", "deviance", "var(dev)"")
for(j in 1:Model.N)
{
m.j <- jags.model(file=model[[j]], data=data.jags[[j]], n.chain=n.chain, n.adapt=100, inits=rngInits[(j-1)*n.chain + (1:n.chain)])
update(m.j, n.BurnIn)
res.j <- jags.samples(m.j, para[[j]], n.Iter, n.Thin)
Sim.DIC <- c(mean(res.j$deviance)+median(res.j$pD), median(res.j$pD), mean(res.j$pD), mean(res.j$deviance), var(res.j$deviance))
}
B.5 Example of hierarchical generalization models in Study 4

_CPT_tr-nm_tr.txt_

# DFD (Task 1): T-K probability weighting and linear value with Trembles model
# DFE-S (Task 2): Natural-mean heuristic with Trembles model

model {

  ##------- Prior Specification -------
  # Priors for individual-level parameters
  for(j in 1:Nsubj[1]) {
    gamma1[j] <- phi(gamma1.phi[j])
    gamma1.phi[j] ~ dnorm(mu.phi.gamma1, tau.phi.gamma1)T(-3, 3)
    eps1[j] <- phi(eps1.phi[j])
    eps1.phi[j] ~ dnorm(mu.phi.eps1, tau.phi.eps1)
  }
  for(j in 1:Nsubj[2]) {
    eps2[j] <- phi(eps2.phi[j])
    eps2.phi[j] ~ dnorm(mu.phi.eps2, tau.phi.eps2)
  }

  # Priors for group-level hyper-parameters
  mu.phi.gamma1 ~ dnorm(0,1)
  tau.phi.gamma1 <- pow(sigma.phi.gamma1,-2)
  sigma.phi.gamma1 ~ dunif(0,10)
  mu.gamma1 <- phi(mu.phi.gamma1)
  mu.phi.eps1 ~ dnorm(0, 1)
  tau.phi.eps1 ~ dgamma(.5, .5)
  mu.eps1 <- phi(mu.phi.eps1)
  mu.phi.eps2 ~ dnorm(0, 1)
  tau.phi.eps2 ~ dgamma(.5, .5)
  mu.eps2 <- phi(mu.phi.eps2)

  ##------- Model Specification and Fitting -------
  for(j in 1:Nsubj[1]) # Subject-loop, DFD
  {
    for(i in 1:Ngam[1]) # Item-loop
    {
      ## Risky prospects
      # Linear value
      v1.rh[i,j] <- prospects.1[i,2,j]
      v1.rl[i,j] <- prospects.1[i,3,j]
# TK Probability weighting function
w1.rh[i,j] <- pow(prospects.1[i,1,j], gamma[j]) / w1.r.z[i,j]
w1.rl[i,j] <- pow((1-prospects.1[i,1,j]), gamma[j]) / w1.r.z[i,j]
w1.r.z[i,j] <- pow(pow(prospects.1[i,1,j], gamma[j]) + pow((1-prospects.1[i,1,j]), gamma[j]), (1/gamma[j]))

## Safe prospects
# Linear value
v1.rm[i,j] <- prospects.1[i,4,j]

## Bernoulli decisions
# Subjective value
sv1.r[i,j] <- w1.rh[i,j] * v1.rh[i,j] + w1.rl[i,j] * v1.rl[i,j]
sv1.s[i,j] <- v1.rm[i,j]

# Choice mechanism: Exponential Luce's rule
dsv.1[i,j] <- step(sv1.r[i,j]-sv1.s[i,j])
binval1[i,j] <- dsv.1[i,j] + (.5-dsv.1[i,j]) * eps1[j]
d1.data[i,j] ~ dbern(binval1[i,j])

} # End Item-loop
} # End Subject-loop, DFD

for(j in 1:N.subj[2]) # Subject-loop, DFE-S
{
    for(i in 1:N.gam[2]) # Item-loop
    {
        # Choice mechanism: Trembles model
        binval2[i,j] <- dnm.2[i,j] + (.5-dnm.2[i,j]) * eps2[j]
        d2.data[i,j] ~ dbern(binval2[i,j])
    } # End Item-loop
} # End Subject-loop, DFE-S
} # End of model