Safe models for risky decisions

Steingröver, H.M.

Chapter 11

Summary and Future Directions

This thesis, entitled "Safe Models for Risky Decisions", scrutinized assumptions about the performance of healthy participants on the Iowa gambling task (IGT; Bechara et al., 1994) and challenged the trustworthiness of conclusions typically obtained from fitting reinforcement-learning models to IGT data. We argued that the risk of drawing premature conclusions from behavioral analyses and computational modeling can be minimized if, in future applications of the IGT and RL models, researchers follow a number of crucial steps. These steps concern behavioral data analyses, model selection, model fitting, and assessment of absolute model fit. In particular, we advocate Bayesian techniques involving Bayesian repeated measures ANOVA for behavioral data analyses, the Bayes factor for model selection, the Bayesian hierarchical framework for model fitting, and posterior predictives to assess the absolute account of the models for the data at hand. Before taking some distance to place the results in broader perspective, I first summarize and discuss the main conclusions of this dissertation.

11.1 Summary of Results

Chapter 2 gave an overview of the performance of healthy participants on the IGT, and pointed to behavioral findings that question key assumptions about IGT performance of healthy participants. Specifically, we showed that (1) healthy participants often fail to develop a pronounced preference for both good decks: Instead, participants often prefer the decks with infrequent losses (i.e., frequency-of-losses effect); (2) healthy participants show idiosyncratic choice behavior (see Figure 11.1); and (3) healthy participants do not show a systematic transition from an initial exploration phase of the decks to an exploitation phase. These findings question the prevailing interpretation of IGT data and suggest that, in future applications of the IGT, key assumptions about performance of healthy participants warrant closer scrutiny.

Chapter 3 used the behavioral findings of Chapter 2, and investigated whether they are in line with the data-fitting potential of three popular reinforcement-learning (RL) models—the Expectancy Valence model (EV; Busemeyer & Stout, 2002), the Prospect Valence Learning model (PVL; Ahn et al., 2008), and a combination of these models, the EV-PU model. However, parameter space partitioning (PSP) revealed important discrepancies between the model-specific popularity and empirical popularity of several choice patterns. In particular, all three models fail to generate pronounced deck preferences often observed in experiments, and the EV model –the first model proposed for the IGT– fails to generate a frequency-of-losses effect that is prominent in healthy participants. This suggests that the EV model should not be used if the data display a

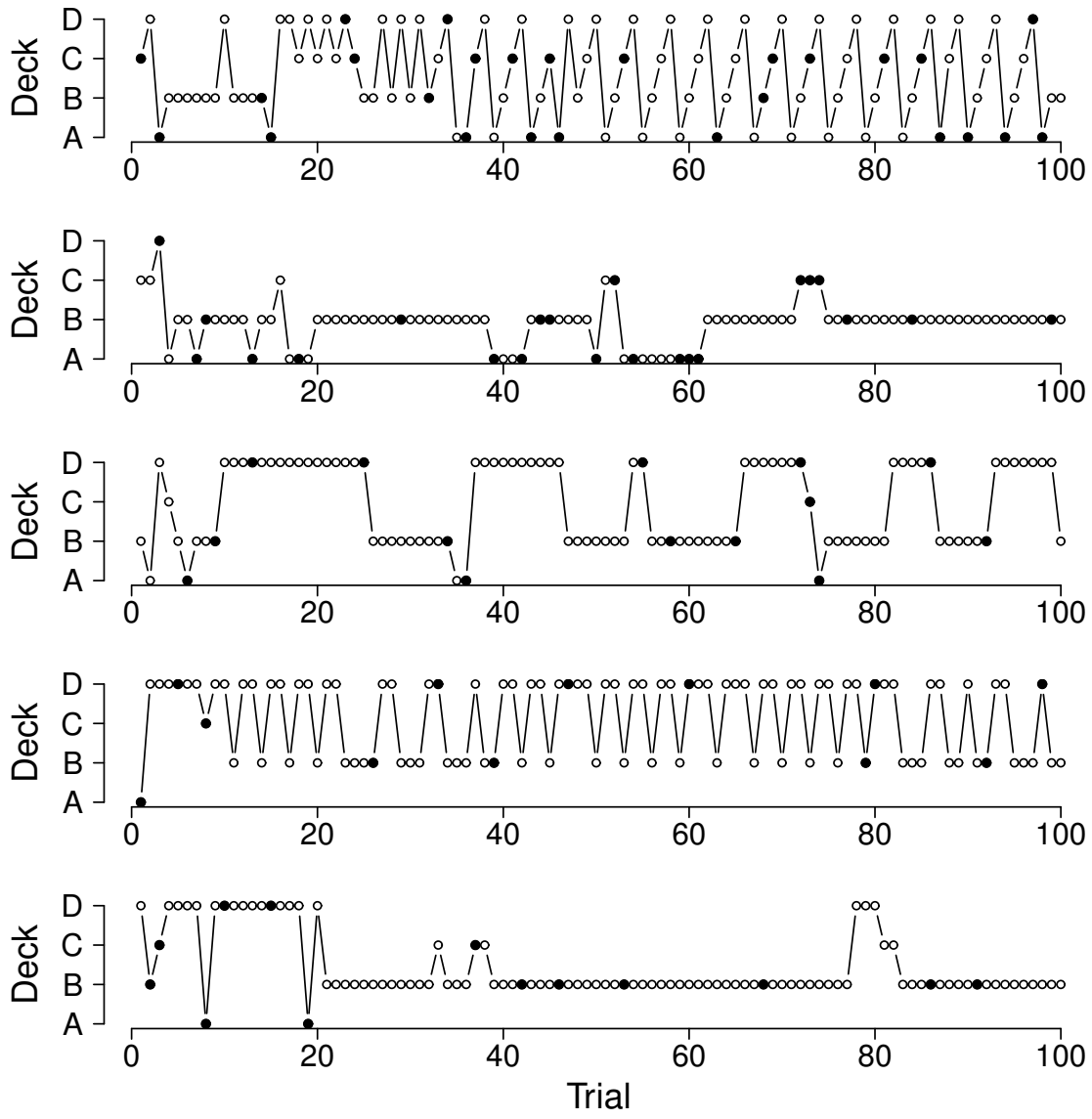

Figure 11.1: Deck selection profiles of five participants express the high within-group variability in the performance of healthy participants. The filled circles indicate the occurrence of rewards and losses together; the empty circles indicate the occurrence of only rewards. The data are published in Steingroever, Wetzels, Horstmann, et al. (2013).

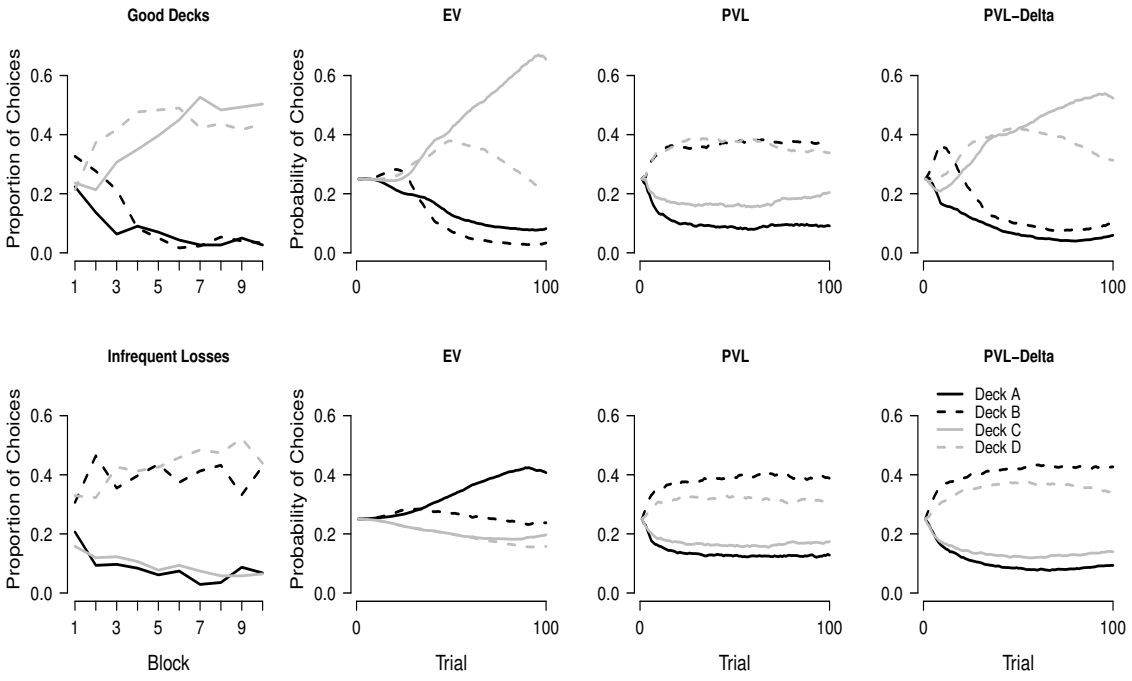

Figure 11.2: Simulation performance of the three RL models with respect to the two stylized data sets used in Chapter 5. The first column presents the observed mean proportions of choices from each deck within 10 blocks. Each block contains 10 trials. The second, third, and fourth column present the mean probabilities of choosing each deck on each trial as generated with the EV, PVL, and PVL-Delta model, respectively.

frequency-of-losses effect; in such a case, the EV model will probably provide a poor fit to the data and thus an inaccurate description of the relevant psychological processes. Taken together, our results of Chapter 3 suggest that the three models under consideration have sufficient data-fitting potential for only a restricted number of choice patterns, and a model that can be universally used is still lacking.

Chapter 4 illustrates three important methods for model validation –parameter recovery, parameter space partitioning, and test of specific influence– as applied to the PVL-Delta model, yet another combination of the EV and PVL models. Our results suggest that the PVL-Delta model (1) recovers parameter accurately; (2) accounts for empirical choice patterns featuring a preference for the good decks or the decks with infrequent losses; (3) fails to account for empirical choice patterns featuring a preference for the bad decks; and (4) performs moderately on the test of selective influence that investigates the effectiveness of experimental manipulations designed to target only a single model parameter. In particular, the test of specific influence showed that the manipulations were successful for all but one parameter. To conclude, despite a few shortcomings, the PVL-Delta model seems to be a better IGT model than the popular EV and PVL models.

Chapter 5 exposes the current practice to draw inferences from parameters of RL models without first having sufficiently assessed the absolute fit of the models for the data. We illustrated the importance of assessing absolute model performance using the EV, PVL, and PVL-Delta models and data from two stylized data sets and five published data sets. Our results showed that all models provided an acceptable fit to the data sets (i.e., post hoc fit); however, when the model parameters were used to generate choices, only the PVL-Delta model captured the qualitative patterns in the

data (i.e., simulation performance; see Figure 11.2). Our results highlight that a model's ability to fit a particular choice pattern does not guarantee that the model can also generate that same choice pattern. However, such ability is crucial to ensure meaningful conclusions from the model parameters. In future applications of the RL models absolute model performance should therefore carefully be assessed to avoid premature conclusions from the model parameters.

Chapter 6 is a rejoinder to Konstantinidis et al. (2014)'s reply on Chapter 5. In this chapter, we stressed that the initial goal of Chapter 5 was *not* to conduct a model selection exercise, but to illustrate why applied researchers should carefully assess absolute model performance before they draw conclusions from the estimated parameters. In addition, we elaborated on the advantages and drawbacks of both the post hoc absolute fit method and the simulation method. A crucial drawback of the post hoc absolute fit method is that in a strict sense this method does not predict, but post-dicts because it uses the data twice; once to fit the model, and once to generate "predictions" which are really postdictions. In addition, we pointed out that model selection criteria based on the post hoc fit method (e.g., the BIC and G^2 criteria) do not fully account for model complexity because they only consider one dimension of model complexity (i.e., the number of model parameters). Finally, we highlighted the distinction between statistical aspects of model adequacy (e.g., good fit to the observed data and good predictions for new data) and psychological relevance of parameter estimates (e.g., good parameter recovery and good performance on tests of selective influence).

Chapter 7 showed how importance sampling can be used to obtain Bayes factors to compare Bayesian individual-level implementations of four RL models using data of 771 healthy participants. In contrast to the BIC and G^2 criteria discussed in Chapter 6, the Bayes factor is a model selection tool that coherently and completely discounts model complexity. Our results provide strong evidence for the Value-Plus-Perseveration (VPP; Worthy, Pang, & Byrne, 2013) model and moderate evidence for the PVL model, but little evidence for the EV and PVL-Delta models (see Figure 11.3). However, we pointed out that it is crucial to interpret our results in combination with results obtained from other model comparison studies in order to obtain a balanced and comprehensive assessment of model adequacy. For instance, poor parameter recovery and simulation performance of the VPP model (Ahn et al., 2014) –an eight-parameter model– suggest that the parameters might have little psychological value, and that a more thorough analysis of the validity of the VPP model is required.

Chapter 8 built on the insights from Chapter 7 and showed how bridge sampling can be used to obtain Bayes factors for both an individual-level Bayesian implementation of the EV model (Figure 11.4) and a hierarchical implementation. We showed that bridge sampling is not only reliable, accurate, and efficient, but also relatively straightforward to implement. In addition, bridge sampling is suitable for models in mathematical psychology that are often complex, non-nested, and hierarchical.

Chapter 9 proposed a suite of three complementary model-based methods for assessing the cognitive variables and processes underlying IGT performance: (1) Bayesian hierarchical parameter estimation; (2) Bayes factor model comparison; and (3) Bayesian latent-mixture modeling. To illustrate these Bayesian analysis techniques, we tested the extent to which differences in decision style (i.e., intuitive, affective vs. deliberate, planned) explain differences in IGT performance. Our results suggest that, on a behavioral level, intuitive and deliberate decision makers behave similarly on the IGT, and the modeling analyses consistently showed that these behavioral preferences are driven by similar cognitive processes. This chapter offered two major methodological advances: (1) the use of the product space method to obtain Bayes factors for comparing two groups of decision makers in a hierarchical Bayesian framework; and (2) the use of a hierarchical mixture model to infer group memberships (Figure 11.5).

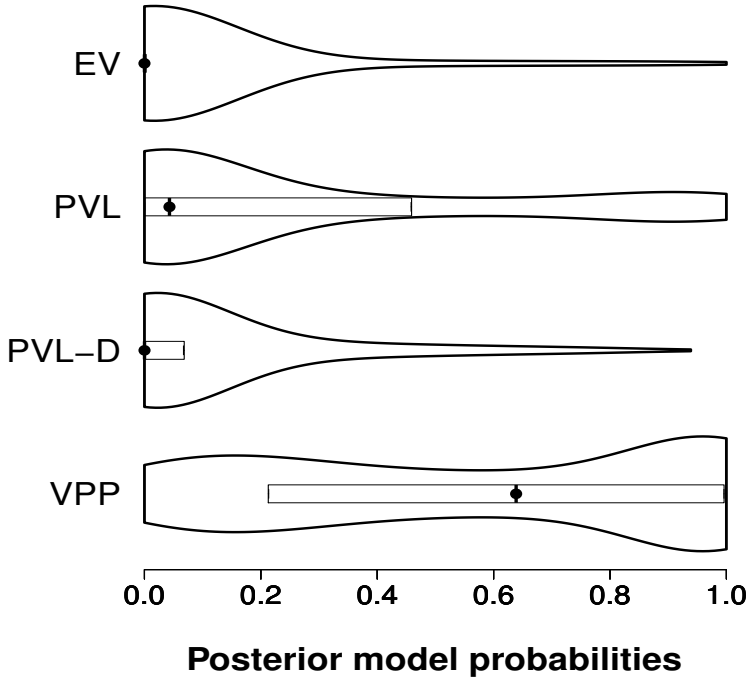

Figure 11.3: Distribution of the posterior model probabilities of 771 participants derived with importance sampling. Each violin plot shows the distribution of posterior model probabilities for one model. The dots indicate the median posterior model probability, and the boxes indicate the interquartile range (i.e., the distance between the .25 and .75 quantiles).

Chapter 10 proposed a Bayesian regression framework that can be used to extend existing Bayesian hierarchical implementations of RL models. This framework allows researchers to obtain Bayes factors in order to quantify the evidential support for relationships between covariates (e.g., decision style) and model parameters. The advantage of this method is that it avoids the multi-step procedure from Chapter 9 where we divided all participants into two groups depending on their score on the decision style covariate, and subsequently tested the groups of participants for differences in their estimated model parameters.

11.2 Future Directions

Several challenges confront researchers who wish to understand the psychological processes that influence risky decision making. These challenges concern (1) the experimental paradigm that is used to investigate how risky decisions are made; (2) the cognitive models that are used to disentangle the driving processes; and (3) the methods that are used to fit and compare the models. Therefore, to advance our understanding of risky decision making several toeholds are possible.

First, the results of Chapter 2 suggest that we need a better paradigm to investigate how risky decisions are made in a controlled, experimental setting (see also Chiu & Lin, 2007; Dunn et al., 2006; Lin et al., 2007). On the one hand, this goal can be achieved by improving the

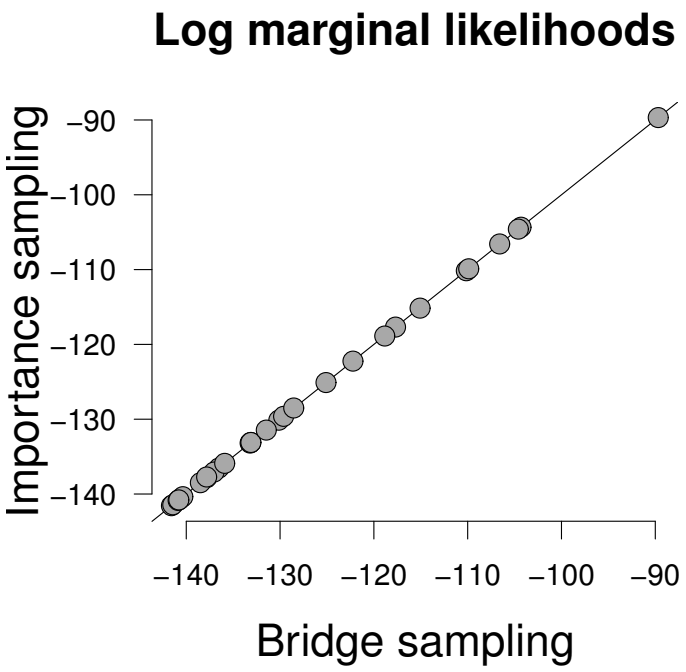

Figure 11.4: Comparison of the log marginal likelihoods of 30 participants of Busemeyer and Stout (2002) obtained with bridge sampling (x-axis) and importance sampling (y-axis) reported in Chapter 7. The solid line expresses perfect correspondence of the two methods.

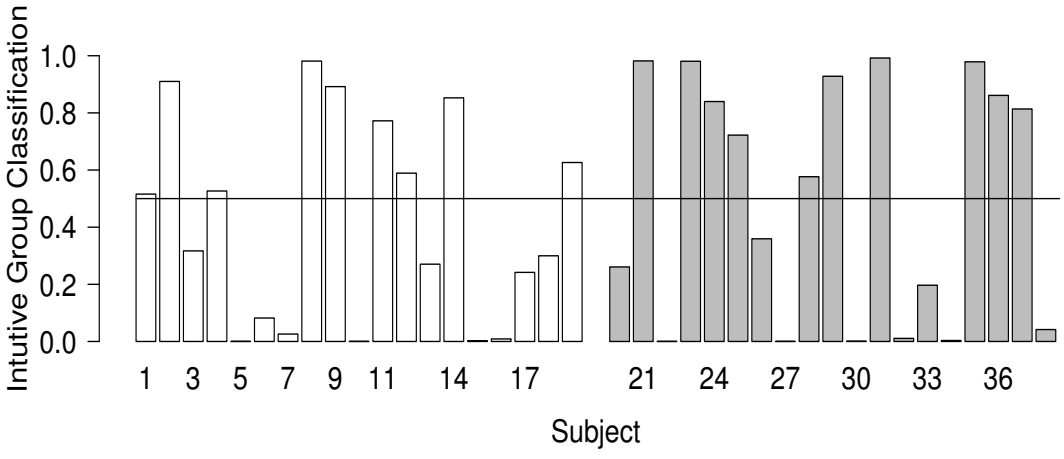

Figure 11.5: Posterior classification as belonging to the group of intuitive decision makers. According to the decision style inventories, participants 1-19 were classified as deliberate decision makers (i.e., white bars), whereas participants 20-38 were classified as intuitive decision makers (i.e., grey bars). The horizontal line represents a posterior classification of .5.

IGT to become more diagnostic and informative, and a valid measure of risky decision making (Buelow & Suhr, 2009). Possibilities are to (1) use at least 150 trials instead of 100 trials (e.g., Wetzels, Vandekerckhove, et al., 2010); (2) avoid the concurrent presentation of wins and losses, but immediately present the net outcome; (3) change the IGT payoff scheme (for several attempts see Appendix B that presents three different IGT payoff schemes, and Chiu & Lin, 2007); and (4) obtain data from additional measures, such as response time (RT) measures, post-decision wagering, confidence ratings for the choices, or quality ratings for the decks. Another approach is to propose a new risky decision-making task that retains the essential characteristics of the IGT (i.e., incorporation of the tradeoff between immediate rewards, but long-term negative consequences vs. safe options that are profitable in the long-run), and at the same time also accounts for the shortcomings illustrated in Chapter 2. Previous research has already pursued this approach, and yielded, for example, the Columbia card task (Figner et al., 2009) and the Soochow gambling task (Lin et al., 2009). Such advances towards a better experimental task to investigate risky decision making –be it a modification of the IGT or a complete new paradigm– are essential, especially given the goal to use such a task for individual diagnoses (Bechara, 2007).

Second, despite the meticulous model development and comparison efforts of the last 15 years, this dissertation underlines that the search for an appropriate IGT model is still subject to ongoing research. Past research focused on developing new models that are a combined version of existing RL models (e.g., the PVL-Delta and EV-PU models) or a slightly adapted variant of existing RL models (e.g., Dai et al., 2015; Pang, Blanco, Maddox, & Worthy, 2016). In addition, past research focused on models that include additional processes, such as perseveration (Worthy, Pang, & Byrne, 2013). Alternatively, to advance the search for an appropriate IGT model, it might be promising to explore models that differ from the traditional RL framework more strongly, such as the instance-based learning model (Gonzalez & Dutt, 2011). Yet, another approach is to enrich current modeling approaches by combining them with models that focus on other processes, such as RT or wagering (Konstantinidis & Shanks, 2014; Persaud, McLeod, & Cowey, 2007).

Third, the state-of-the-art framework for parameter inference can be improved even further. We generally advocated the Bayesian hierarchical framework; however, the idiosyncratic choice behavior of healthy participants reported in Chapter 2 suggests that it may be a mistake to assume a single group of healthy participants. Instead, it might be more realistic to use a Bayesian hierarchical framework that contains a mixture component (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014; Huizenga et al., 2007, and Chapter 9). In addition, parameter inference can be improved by developing more informative prior distributions based on published empirical findings (e.g., Gershman, 2016).

Moreover, it is important to note that the results of this dissertation, but also many other model comparison studies over the last 15 years, are mainly based on data of healthy participants. It is therefore informative and importing to investigate whether our results generalize to clinical populations. The RL models are often applied to IGT data of clinical groups to better understand their decision-making deficits, and therefore it is crucial that conclusions from model parameters are trustworthy. Hence we recommend that absolute model performance is assessed also for clinical datasets. In addition, we recommend that sophisticated model comparison tools are used to identify a model that can jointly capture psychological processes driving IGT performance of clinical and healthy participants. Moreover, it is desirable to test the assumptions underlying IGT performance of patients with lesions to the ventromedial prefrontal cortex (vmPFC) analogous to the procedure reported in Chapter 2 focusing on healthy participants. Such research could reveal whether Bechara et al. (1994)'s key assumptions at least hold for the clinical population of vmPFC patients. Yet another insightful project would be to try and replicate Yechiam et al. (2005)'s mapping of the model parameters of 10 different clinical groups using state-of-the-art methods.

Last but not least, it is important to be aware that modeling IGT data is of interest to a diverse group of researches and practitioners (e.g., clinical psychologists, statisticians, mathematical psychologists, and computer scientists). To advance our understanding of risky decision making, it is therefore important to further collaborations between these different groups (Ahn, Dai, Vassileva, Busemeyer, & Stout, 2016). In addition, it is important to make model-fitting routines easier available and applicable (e.g., by adding them to JASP; JASP Team, 2015; or by providing an R package; Ahn, Haines, & Zhang, 2016). Clear guidelines and support are needed to facilitate the choice of an appropriate IGT model out of the large available pool of different models, to explain how the to fit the models, how to assess the account of the models for the data, and how to interpret the results. Exemplary steps towards this ambition have already been initiated by Ahn, Haines, and Zhang (2016).

11.3 Concluding Remarks

The development of the Iowa gambling task (Bechara et al., 1994) and the application of reinforcement-learning models greatly advanced research efforts about risky decision making. On the one hand, these research effort led to many achievements as pointed out by Ahn and Busemeyer (2016): "computational modeling has greatly contributed to understanding cognitive processes underlying our decision-making" (p. 1; see also the special issue "Iowa Gambling Task (IGT): Twenty Years After" in Frontiers in Psychology). Another summary of the achievements is given by Ahn, Dai, et al. (2016, p. 62):

> The past 10 years have seen the development of cognitive models for the IGT and SGT, with adequate model fits, and parameters that appear to have good utility for distinguishing between various clinical samples, and that relate to significant individual characteristics such as personality measures and severity of clinical symptoms. These have deepened the understanding of the variety and nature of differences between various substance abuse and other clinical groups, opening a potential window into the way basic psychological processes such as learning from experience or feedback, and sensitivity to reward and punishment, may be affected by substance abuse or may create vulnerability factors for developing substance use disorders. Furthermore, these models have provided a possible way in which individual characteristics can be assessed and targeted in individually tailored treatments.

On the other hand, there is still much progress ahead, as pointed out in this dissertation and also by several other researchers; Ahn and Busemeyer (2016), for example, admit that "Despite the growing enthusiasm, no computational assays or methods have influenced clinical practice yet" (p. 1). In addition, Gershman (2016) point to problems of RL models by stating that "fitting the parameters of these models can be challenging: the parameters are not identifiable, estimates are unreliable, and the fitted models may not have good predictive validity" (p. 1).

Discussing a large variety of models and methods to compare the models, this dissertation illustrated that research efforts about risky decision making greatly advanced during the last years. On the other hand, this dissertation also illustrated the major challenges by pointing to problems with respect to behavioral analyses and cognitive modeling. In particular, I tried to point out how we can achieve more meaningful conclusions about the psychological processes driving risky decision making. First of all, I argued that it is crucial to change the way how IGT data are typically analyzed. I proposed using Bayesian repeated measures ANOVA, analyses of deck selection profiles, and consideration of group-choices as a function of blocks for each deck

separately to avoid hiding a frequency-of-losses effect and to better reveal the individual deck preferences. Second, I argued that it is crucial to assess absolute model performance carefully, using post hoc fit and simulation performance in order to avoid premature conclusions from the model parameters. Third, I illustrated a number of different methods that can be used to compare RL models or the model parameters of different groups of decision makers. Pursuing these suggestions will hopefully advance us towards the ambitious goal of using the IGT –or an alternative risky decision-making paradigm– and cognitive models to reliably measure risky decision making and to understand the underlying psychological processes.