Safe models for risky decisions

Steingröver, H.M.

Chapter 9

Bayesian Techniques for Analyzing Group Differences in the Iowa Gambling Task: A Case Study of Intuitive and Deliberate Decision Makers

This chapter has been submitted for publication as:
Helen Steingroever, Thorsten Pachur, Martin Šmíra, and Michael D. Lee (2016). Bayesian techniques for analyzing group differences in the Iowa gambling task: A case study of intuitive and deliberate decision makers.

Abstract

The Iowa gambling task (IGT) is one of the most popular experimental paradigms for assessing real-world decision making. In order to understand the psychological processes that underlie IGT performance, several cognitive models have been developed. For comparing how groups of individuals make decisions on the IGT, the standard approach is to estimate the maximum likelihood model parameters for each person, and then conduct frequentist tests to compare the parameters across groups. Here, we present a Bayesian alternative. In addition to a Bayesian repeated measures ANOVA for comparing behavioral performance, we propose 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 test 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 show that these behavioral preferences are driven by similar cognitive processes. Thus, our results challenge the notion that individual differences in intuitive and deliberative decision styles have a very broad impact on decision making, and that intuitive processes in healthy adults play a central role for IGT performance. The ability to draw such a conclusion demonstrates a major advantage of the Bayesian approach, that is, to be able to quantify evidence in favor of the null hypothesis.

The Iowa gambling task (IGT; Bechara et al., 1994)—a card game that consists of two bad decks (i.e., negative long-term outcomes) and two good decks (i.e., positive long-term outcomes)—is arguably the most popular neuropsychological paradigm for assessing decision-making deficits in clinical populations (Toplak et al., 2010). There is considerable evidence that the IGT performance of healthy decision makers (i.e., participants that do not have any neurological impairments) differs from that of clinical populations such as patients with lesions to the ventromedial prefrontal cortex (Bechara et al., 1998, 1999, 2000), pathological gambling (Cavedini, Riboldi, Keller, et al., 2002), obsessive-compulsive disorder (Cavedini, Riboldi, D'Annucci, et al., 2002), psychopathic tendencies (Blair et al., 2001), or schizophrenia (Bark et al., 2005; Martino et al., 2007).

These studies have mainly relied on an analysis of the proportion of choices from the good decks as compared to the bad decks, with subsequent conclusions about group differences based on frequentist analysis techniques, such as t-tests and analyses of variance (ANOVAs). In addition, to investigate whether two groups differ in the psychological processes that underlie their performance, several reinforcement-learning (RL) models have been proposed. These models assume that card selection on the IGT results from an interaction of distinct psychological processes such as motivation, memory, and response consistency (Busemeyer et al., 2003). Using these models, it has been possible to infer group differences in cognitive processes even when the behavior of the groups at the aggregate level does not differ (e.g., Yechiam, Kanz, et al., 2008). Popular RL models for IGT data are the Expectancy Valence model (EV; Busemeyer & Stout, 2002; Yechiam, Kanz, et al., 2008) and the Prospect Valence Learning model (PVL; Ahn et al., 2008, 2011; see Steingroever, Wetzels, & Wagenmakers, 2013a, for additional references and a detailed description of the EV and PVL models). More recently, it has been shown that the hybrid PVL-Delta model outperforms the EV and PVL model in many model comparison analyses (Ahn et al., 2008; Fridberg et al., 2010; Steingroever, Wetzels, & Wagenmakers, 2013b; Steingroever et al., 2014, but see also Worthy, Pang, & Byrne, 2013; for the Value-Plus-Perseveration model).

The current standard approach for comparing model parameters between groups is to estimate the parameters for each participant separately using maximum likelihood, and then to use frequentist statistical tests, such as independent-samples t-tests, Friedman tests, or Mann-Whitney U tests, to compare the estimates across groups (e.g., Cella et al., 2012; Escartin et al., 2012; Yechiam, Kanz, et al., 2008; Yechiam, Hayden, et al., 2008). However, individual-level maximum likelihood results in inferior parameter inferences compared to Bayesian hierarchical parameter estimation (Ahn et al., 2011; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, et al., 2010). In addition, there are several well-known problems inherent with frequentist tests, such as p-values overstating the evidence against the null hypothesis (Berger & Delampady, 1987; Edwards, Lindman, & Savage, 1963; V. E. Johnson, 2013; Sellke, Bayarri, & Berger, 2001), and that classical hypothesis testing cannot be used to quantify evidence in favor of the null hypothesis. The latter is a crucial disadvantage, given that many theories predict the absence of an effect (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Finally, in contrast to Bayesian sequential testing, frequentist sequential testing is much less flexible, since it requires researchers to specify in advance the total duration of the data collection period (e.g., Reboussin, DeMets, Kim, & Lan, 2000), and the number of interim analyses (e.g., Pocock, 1977).

Here, we present a Bayesian approach to examine whether two groups differ in their IGT performance, encompassing both behavioral and model-based analyses. We illustrate the Bayesian approach by comparing IGT performance of decision makers with an intuitive (affective) decision style to those with a deliberate (planned) decision style, distinguished via established self-report instruments to measure decision style. This comparison is of theoretical interest because the prominent somatic marker hypothesis (Bechara et al., 1997; A. R. Damasio, Tranel, & Damasio, 1991; A. Damasio, 1994) suggests that intuitive processes are of particular importance for successful

performance on the IGT. We apply a Bayesian repeated measurement ANOVA to investigate whether intuitive and deliberate decision makers differ in their deck preferences, and illustrate three complementary cognitive analyses for comparing the groups on parameters estimated with the PVL-Delta model: (1) Bayesian hierarchical parameter estimation, (2) Bayes factor model comparison, and (3) Bayesian latent-mixture modeling (see also Lee, Lodewyckx, & Wagenmakers, 2015). All our analyses were conducted using JASP (JASP Team, 2015), R (R Core Team, 2015), and the Stan software (Stan Development Team, 2014b, 2014c; Hoffman & Gelman, 2014), all of which are freely available. The relevant R and Stan code is available online, and can be adapted for similar IGT models and similar decision-making tasks.

The outline of this article is as follows. The next section describes the IGT, the PVL-Delta model, and its Bayesian hierarchical implementation together with a brief review of Bayesian statistics. The following sections present the proposed methodology, and then its application to IGT data from intuitive and deliberate decision makers. In the final section, we summarize our findings and discuss the methodological contribution of our proposed analysis approach, as well as implications for the debate on the role of intuition and deliberation in the IGT.

9.1 The IGT and PVL-Delta Model

The IGT

In this section we describe the IGT (see also Steingroever, Wetzels, Horstmann, et al., 2013; Steingroever, Wetzels, & Wagenmakers, 2013a, 2013b; Steingroever et al., 2014, 2016). In the traditional version of the task, participants are initially given $2000 (hypothetically) and are presented with four decks of cards with different payoffs. Participants are instructed to choose, over several rounds, cards in order to maximize their long-term net outcome (Bechara et al., 1994, 1997). Unbeknownst to the participants, the task (typically) contains 100 trials. After each choice, participants receive feedback on the rewards and losses (if any) associated with that card, as well as their running tally of rewards and losses over all trials so far.

Table 9.1: *Summary of the payoff scheme of the traditional IGT as developed by Bechara et al. (1994).*

	Deck A	Deck B	Deck C	Deck D
	Bad deck with frequent losses	Bad deck with infrequent losses	Good deck with frequent losses	Good deck with infrequent losses
Reward/trial	100	100	50	50
Number of losses/10 cards	5	1	5	1
Loss/10 cards	−1250	−1250	−250	−250
Net outcome/10 cards	−250	−250	250	250

A crucial aspect of the IGT is whether and to what extent participants eventually learn to prefer the good decks because only choosing from the good decks maximizes their long-term net outcome. The good decks are typically labeled as decks C and D, whereas the bad decks are labeled as decks A and B. Table 9.1 presents a summary of the traditional payoff scheme as developed by Bechara et al. (1994). This table illustrates that decks A and B yield high constant rewards, but even higher unpredictable losses: hence, the long-term net outcome is negative. Decks C and D, on the other hand, yield low constant rewards, but even lower unpredictable losses: hence, the

long-term net outcome is positive. In addition to the different payoff magnitudes, the decks also differ in the frequency of losses: decks A and C yield frequent losses, while decks C and D yield infrequent losses.

The PVL-Delta Model

In this section, we describe the PVL-Delta model (see also Steingroever, Wetzels, & Wagenmakers, 2013b; Steingroever et al., 2016). The model formalizes participants' performance on the IGT through the interaction of four parameters that have natural psychological interpretations as controlling different psychological processes (Ahn et al., 2008; Fridberg et al., 2010; Steingroever et al., 2014).

The first assumption of the PVL-Delta model is that, after choosing a card from deck $k \in \{1, 2, 3, 4\}$ on trial t, participants evaluate the net outcome associated with the card using prospect theory (Tversky & Kahneman, 1992). Formally, the utility is given by

$$u_k(t) = \begin{cases} X(t)^A & \text{if } X(t) \geq 0 \\ -w \cdot |X(t)|^A & \text{if } X(t) < 0. \end{cases} \tag{9.1}$$

In this equation, $X(t)$ represents the net outcome on trial t, which is the sum of the experienced reward and loss (i.e., $X(t) = W(t) - |L(t)|$). The prospect utility function contains the first two model parameters. These are the loss aversion parameter $w \in [0, 5]$, and the outcome sensitivity parameter $A \in [0, 1]$.

The loss aversion parameter w quantifies the relative weight of net losses relative to net gains in participants' evaluation of the net outcome of a given card. A value of w greater than one indicates a larger impact of negative than of positive net outcomes, whereas a value of w approaching one indicates a similar impact of negative and positive outcomes. As w approaches zero, the model predicts that negative net outcomes will be neglected.

The outcome sensitivity parameter A quantifies the extent to which the subjective utility corresponds to the actual value, $X(t)$. As A approaches one, the subjective utility $u_k(t)$ increases in proportion to the actual net outcome. For values of A smaller than one, there is less differentiation between the positive and negative net outcomes. As A approaches zero, the sensitivity to differences in the positive and negative net outcomes continues to decrease towards the limit in which there is no sensitivity.

The PVL-Delta model also assumes that, having formed the utility of the card through Equation 9.1, people update their expected utility of the just-chosen deck, but keep the expected utilities of the remaining decks unchanged. This updating process is described by the delta learning rule:

$$Ev_k(t) = Ev_k(t-1) + a \cdot (u_k(t) - Ev_k(t-1)). \tag{9.2}$$

The delta learning rule states that the expected utility of the chosen deck k is adjusted upward if the experienced utility $u_k(t)$ is higher than expected. If the experienced utility $u_k(t)$ is lower than expected, the expected utility of deck k is adjusted downward.[1] This updating process is influenced by an updating parameter $a \in [0, 1]$. This parameter expresses the memory for past expectancies. A value of a close to zero indicates slow forgetting and weak recency effects, whereas a value of a close to one indicates rapid forgetting and strong recency effects.

In the next step, the PVL-Delta model assumes that the expected utilities of each deck guide participants' choices on the next trial. This assumption is formalized by the softmax choice rule,

[1] We initialized the expectancies of each deck k to zero so that $Ev_k(0) = 0 \; \forall k$.

also known as the ratio-of-strength choice rule (Luce, 1959). The PVL-Delta model uses this rule to compute the probability of choosing each deck on each trial (Equation 9.3). The softmax choice rule includes a sensitivity parameter θ that controls the extent to which trial-by-trial choices match the expected deck utilities. Values of θ close to zero indicate random choice behavior (i.e., strong exploration), whereas large values of θ indicate choice behavior that is strongly determined by the expected utilities (i.e., choices strictly follow the expectancies of the decks).

$$P[S_k(t+1)] = \frac{e^{\theta \cdot Ev_k(t)}}{\sum_{j=1}^{4} e^{\theta \cdot Ev_j(t)}} \tag{9.3}$$

The PVL-Delta model assumes a trial-independent sensitivity parameter θ, which depends on the final model parameter: the response consistency $c \in [0, 5]$ (Equation 9.4). Small values of c lead to a small values of sensitivity θ and thus to more random choice patterns, whereas large values of c lead to larger values of θ, and thus to more deterministic choice patterns.

$$\theta = 3^c - 1 \tag{9.4}$$

In sum, the PVL-Delta model has four parameters: (1) an outcome sensitivity parameter A, which determines the shape of the utility function, (2) a loss aversion parameter w, which quantifies the weight of net losses over net rewards, (3) an updating parameter a, which determines the memory for past expectancies, and (4) a response consistency parameter c, which determines the balance between exploration and exploitation in the deck choices.

Bayesian Hierarchical Implementation of the PVL-Delta Model

We used a Bayesian hierarchical implementation of the PVL-Delta model for the three cognitive modeling analyses (see Steingroever, Wetzels, & Wagenmakers, 2013b, for more details on the implementation, and Wetzels, Vandekerckhove, et al., 2010, for the same model specification in the case of the EV model). In a Bayesian hierarchical framework, the parameters of individual subjects of a specific group are assumed to stem from a group-level distribution. The Bayesian hierarchical framework thus naturally incorporates both the differences and commonalities between and within the participants of one group, and produces both inferences about individual-level and group-level parameter (Horn et al., 2015; Lejarraga et al., 2016; Navarro et al., 2006; Rouder & Lu, 2005; Rouder et al., 2005, 2008). To confirm that we correctly implemented the PVL-Delta model, we ran several parameter-recovery studies. The results of two such studies are presented in Appendix G.

In Bayesian parameter estimation, inferences about a parameter are based on the posterior distribution of the parameter values given the observed data. A posterior distribution expresses the uncertainty about the value of a parameter based on the modeling assumptions and the observed data.

In the Bayesian framework, Bayes factors are used to choose between models and to test hypotheses (Berger & Mortera, 1999; Edwards et al., 1963; Jeffreys, 1961; Kass & Raftery, 1995; Rouder, Morey, Speckman, & Province, 2012; Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers et al., 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). The Bayes factor quantifies the relative probability of the data under two competing models or hypotheses. In particular, BF_{01} quantifies the probability of the data under the null hypothesis (H_0) relative to the probability of the data under the alternative hypothesis (H_1). A Bayes factor can, for example, be used to quantify the evidence that the data provide for a model that assumes differences in the loss aversion parameter across two groups of decision makers ($\mathcal{M}_1$), compared to a model that

assumes no differences ($\mathcal{M}_0$). If, for example, it was found that $\text{BF}_{01} = 10$, this would indicate that the data were 10 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$. Alternatively, if it were found that $\text{BF}_{01} = 1/10$, or equivalently, that $\text{BF}_{10} = 10$, this would indicate that the data were 10 times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$. As these possibilities make clear, Bayes factors, in contrast to frequentist methods, allow the evidence for the null hypothesis or null model to be quantified (e.g., Rouder et al., 2009).

9.2 Proposed Methodology for Comparing Groups on the IGT

The IGT has often been used to investigate group differences in decision making. Group differences are interesting because the IGT is assumed to tap into a broad spectrum of distinct psychological processes; by comparing group differences in performance—and in particular by decomposing the behavior using cognitive modeling—there is the potential to identify which processes are different and which are the same. For example, Yechiam and colleagues found that drug and sex offenders over-weighted potential gains as compared with losses, whereas assault criminals tended to make less consistent choices and to focus on immediate outcomes (Yechiam, Kanz, et al., 2008). These findings required the use of a cognitive model because basic data analyses of the card selection behavior did not show substantial differences.

Accordingly, in this section, we present a set of Bayesian statistical analyses that can be applied to compare the performance of groups of people on the IGT. We start with a standard method for behavioral data analysis, before proposing a more novel set of complementary approaches for applying cognitive models.

Behavioral Data Analyses

Basic behavioral data analyses are usually based on general linear models. A standard IGT experiment involves repeated measures for a number of participants in two or more groups over two or more blocks of trials. Accordingly, a Bayesian block x group repeated measures ANOVA on the choices from the good decks (i.e., decks C and D) is appropriate. These sorts of ANOVA analyses can be conveniently performed in JASP (JASP Team, 2015; Rouder et al., 2012), which is user-friendly free software with a graphical user interface for conducting Bayesian data analysis.

Cognitive Modeling Analyses

We implemented all of our proposed model-based analyses using Stan (Stan Development Team, 2014b, 2014a; Hoffman & Gelman, 2014; see Chapter 9 of Stan Development Team, 2014c, for a description on how to implement mixture models in Stan).

Bayesian hierarchical parameter estimation

The first model-based analysis involves inferring the posterior distributions of the group-level mean parameters across the two groups. These inferences can be made using the Bayesian hierarchical implementation of the PVL-Delta model described earlier, which assumes that the model parameters of each participant are drawn from a group-level distribution (Steingroever, Wetzels, & Wagenmakers, 2013b). To assess the account of the PVL-Delta model to the data we used the post hoc fit method as described in Steingroever et al. (2014).

Bayes factor model comparison

The second model-based analysis involves comparing the group-level mean parameters. This can be achieved by comparing models that assume differences in the group-level mean parameters across the two groups to a model that assumes no differences in these parameters (i.e., a null model). For models like the PVL-Delta that have more than one parameter of interest, multiple comparisons of this type are needed.

When we refer to a model that assumes differences in at least one group-level mean parameter, we index $\mathcal{M}$ by the corresponding group-level mean parameter. $\mathcal{M}_{\mu_w \mu_c}$, for example, refers to the model that assumes differences in the group-level mean parameter of the loss aversion parameter w and of the consistency parameter c (i.e., $\mu_{w,1} \neq \mu_{w,2}$ and $\mu_{c,1} \neq \mu_{c,2}$, where the second index refers to the group), but no differences in group-level mean parameter of the outcome sensitivity parameter A and of the updating parameter a (i.e., $\mu_{A,1} = \mu_{A,2}$ and $\mu_{a,1} = \mu_{a,2}$).

Since the PVL-Delta model has four parameters, we compared a total of $2^4 = 16$ models. These 16 models represent all possible combinations of the four group-level mean parameters of the PVL-Delta model that can either be the same or differ across the two groups. Note that, for all of the model comparisons, we assumed that the group-level standard deviations are the same across the two groups (i.e., $\sigma_{A,1} = \sigma_{A,2}$, $\sigma_{w,1} = \sigma_{w,2}$, $\sigma_{a,1} = \sigma_{a,2}$, and $\sigma_{c,1} = \sigma_{c,2}$). To quantify the evidence that the data provide for each of the 16 models, we used Bayes factors under the assumption of equal prior model probabilities of all models. Due to this assumption, the Bayes factor BF_{01} simplifies to the posterior model odds, that is, the ratio of the posterior probability of model $\mathcal{M}_0$ relative to the posterior probability of model $\mathcal{M}_1$. The posterior probability of a specific model $\mathcal{M}$ was estimated by means of the product space method (Carlin & Chib, 1995; Lodewyckx et al., 2011; see Appendix G for more details on the product space method, and see Steingroever et al., 2016, for more details on Bayes factors).

To confirm the stability of the Bayes factor estimates, we undertook several tests. First, we confirmed good sampling behavior of the model indicator variable z (i.e., good mixing and low autocorrelations, that is, frequent model switches; Lodewyckx et al., 2011). Secondly, we repeated the product space method with fewer iterations (i.e., 5,000 samples instead of 7,000 of each chain after having discarded the first 1,000 samples of each chain as burn-in). The stability of the Bayes factor estimates was confirmed because the difference in corresponding estimated posterior model probabilities was smaller than 0.01. Thirdly, the correctness of our Stan model file was discussed on the Stan users mailing list.[2]

Latent-mixture modeling

The first two model-based analyses focus on parameter estimation and model selection, respectively. Though relatively standard approaches in the general Bayesian statistics literature, they are the exception in the context of the IGT and associated cognitive modeling. The third model-based analysis, which combines elements of parameter estimation and model selection in a complementary way, is novel both in the context of the IGT and in Bayesian applications more generally. This analysis involves a two-group latent hierarchical mixture model (Lee et al., 2015; Chapter 6 in Lee & Wagenmakers, 2013).

The goal of the latent-mixture analysis is to infer the group membership of each participant, based on the cognitive model and the data, but without knowledge of the group membership of each participant. Instead, the group membership of each participant is represented by a latent

[2]The discussion can be found here `https://groups.google.com/forum/?hl=cs#!searchin/stan-users/reinforcemen/stan-users/TjY3wQqUS2g/cff2lWoRUr0J`.

indicator variable, and the analysis infers each participant's probability of belonging to either of the two groups.

Formally, in the two-group case, group membership is indexed by a binary indicator variable z_i (i.e., $z_i = 0$ and $z_i = 1$ indicate that the i-th participant belongs to the first and second group, respectively). The prior for these indicator parameters is $z_i \sim \text{Bernoulli}(\psi)$, so that $\psi \sim \text{Uniform}(0, 1)$ corresponds to the base-rate of membership to the second group. This choice of priors means that each participant is *a priori* equally likely to be assigned to either group. The latent-mixture model analysis provides a posterior distribution of the base-rate, and, for all participants separately, the probability with which they are assigned to each of the groups.

One way to apply this latent-mixture analysis is to use the same priors for model parameters as used in the first analysis to make group-level inferences. In this case, the inferences made by the latent-mixture analysis about the group membership of each participant reflect how people would be classified without knowing the true memberships. If these inferred group memberships agree with the actual one, the analysis provides strong evidence that the behavioral data and model separate people into the proposed groups.

A second way to apply the latent-mixture model approach is to use highly informative priors from the first analysis, so that each group is defined in terms of group-level parameter inferences based on the true memberships. That is, we assumed that probit transformed individual-level parameters were drawn from a group-level normal distribution: $z_i' \sim \text{N}(\mu_{z'}, \sigma_{z'})$. Note that we use z_i to refer to a specific individual-level parameter of the PVL-Delta model (i.e., $z_i \in \{A_i, w_i, a_i, c_i\}$), and z_i' to refer to its probit transformed version (i.e., $z_i' = \Phi^{-1}(z_i)$ with Φ^{-1} being the inverse of the cumulative standard normal distribution function). We assigned a normal prior to the group-level means $\mu_{z'}$, and a truncated normal prior (allowing for only positive values) to the group-level standard deviations, $\sigma_{z'}$. These prior distributions were characterized by means and standard deviations obtained from the first analysis. That is, we used the mean and the standard deviation of the posterior distribution of $\mu_{z'}$ obtained from the first analysis to specify the prior distribution on $\mu_{z'}$ in the informed latent-mixture model approach, and analogous for the prior distribution on $\sigma_{z'}$. This analysis obviously re-uses the behavioral data, and so cannot be used to make inferences about model parameters. It does, however, potentially provide a strong test of patterns of group membership. In particular, if the true group memberships of participants cannot be inferred under these ideal conditions, there is strong evidence that the model and data do not distinguish the participants into the proposed groups.

9.3 Case Study: Intuitive versus Deliberate Decision Making

Whereas many early applications of the IGT focused on comparing clinical to control groups, the task has increasingly also been used to study how individual differences in cognitive abilities (i.e., executive functions, intelligence), state mood, or personality characteristics among healthy participants can explain differences in decision making (Buelow & Suhr, 2009; Suhr & Tsanadis, 2007; Toplak et al., 2010). One interesting variable whose impact on decision behavior has recently received much attention is *decision style* (e.g., Phillips, Fletcher, Marks, & Hine, 2016), which measures whether people prefer making decisions using an intuitive or a deliberate decision mode (Betsch, 2004).

Thus, we considered possible IGT differences between intuitive and deliberative decision makers as a case study for our proposed methodology. The IGT seems a particularly promising context for this purpose because Bechara et al. (1997) proposed that intuitive, affective processes are important for good performance on this task. Even though Betsch (2004) showed that intuitive

and deliberative decision styles are linked to behavioral differences in many decision tasks, such as the valuation of consumer items and monetary lotteries (Schunk & Betsch, 2006; Betsch & Kunz, 2008), we believe ours is the first study that investigates whether decision style also impacts IGT performance. Different predictions about the nature of an effect are conceivable. Based on the somatic maker hypothesis it is reasonable to expect that affective, intuitive processes lead to successful IGT performance and that intuitive and deliberate decision makers might thus also differ with regard to the processes underlying their card selection (Bechara et al., 1997; A. R. Damasio et al., 1991; A. Damasio, 1994). The results presented by Maia and McClelland (2004), however, suggest that consciously accessible, deliberate processes are crucial for good IGT performance.

Data

Seventy students from the University of Basel (average age 24.9, $SD = 5.8$, range $= 19 - 51$ years, 49 female) participated in the study. To measure individual participants' decision style, we used a self-report inventory complied by Betsch and Iannello (2010) consisting of 12 subscales. Based on the mean score for each participant on each subscale, we conducted a principal component analysis with a rotation based on the varimax method. The Kaiser criterion suggested a three-factor solution (i.e., a deliberation factor, an intuition factor, and a spontaneity factor). Following previous research (Betsch & Kunz, 2008), we classified participants as intuitive if they had both a factor score above the median of the intuition factor and a factor score below the median of the deliberation factor. Participants with the opposite pattern were classified as deliberate. This classification scheme yielded 19 participants in the intuitive group and 19 participants in the deliberate group. Thirty two participants thus remained unclassified. Using an alternative classification, which included all participants, and simply distinguished between intuitive and deliberate decision makers based on whether or not their score on the intuition factor was higher than the median, resulted in qualitatively identical conclusions. Appendix G provides more details.

Behavioral Data Analyses

In order to obtain a visual impression of the group-level deck preferences across trials, Figure 9.1 shows the proportion of choices from each deck as a function of 10 blocks, and the proportion of choices from the good and bad decks, separately for intuitive and deliberate decision makers. The figure suggests similar deck preferences for both groups. Specifically, although both groups failed to develop a clear avoidance of bad deck B, overall they learned to make more choices from the good decks than from the bad decks. There appears to be a slight trend for stronger learning in the group of intuitive decision makers.

We applied our proposed Bayesian data analysis, in the form of a 10 (block) x 2 (decision style) repeated measures ANOVA. The results of this analysis showed that the data are 3.64 times more likely under the null model that assumes no group differences in the number of choices from the good decks than under the alternate model that does assume group differences (i.e., the Bayes factor is 3.64 in favor of the model that includes no main effect of group). According to the classification scheme of Jeffreys (1961), this can be considered as moderate evidence for the null model. In addition, the data are about five times more likely under the model that assumes that there is no interaction between block and decision style than under the model that assumes that there is such an interaction effect (i.e., the Bayes factor is 5.38 in favor of the model that includes no interaction effect between block and decision style).[3] This can also be classified as moderate

[3]The frequentist repeated measures ANOVA revealed that neither the main effect of decision style ($F(1, 36) = .404, p = .529$) nor the interaction between block and decision style ($F(9, 324) = 1.466, p = .159$) was significant.

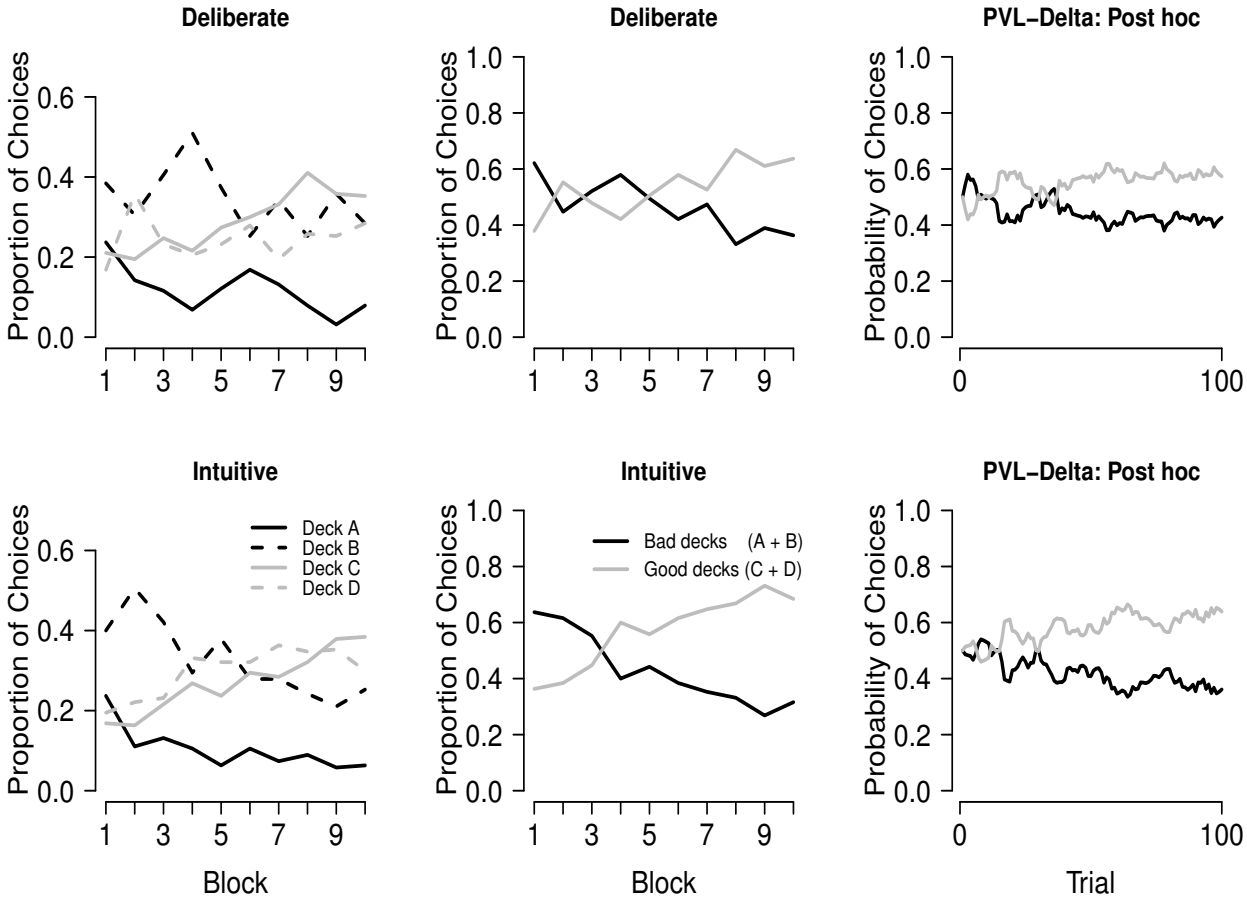

Figure 9.1: Mean proportion of choices from each deck within 10 blocks of both groups of decision makers (first column). Each block contains 10 trials. The second column shows the mean proportion of choices from the good and bad decks within 10 blocks of both groups of decision makers. The third column shows the predictions of the PVL-Delta model for both groups of decision makers. The predictions were obtained by computing the mean probabilities of choosing the good decks or the bad decks on each trial according the post hoc absolute fit method (see Steingroever et al., 2014).

evidence for the null model (Jeffreys, 1961). These results suggest that deliberate and intuitive decision makers show similar learning curves on the IGT.

Cognitive Modeling Analyses

Even though the behavioral data analysis suggests that intuitive and deliberate decision makers show similar deck preferences on the IGT, it might still be the case that there are differences in the underlying cognitive processes driving the decisions of the two groups. To investigate this possibility, we next decompose the IGT performance of the two groups using three different cognitive modeling analyses.

In each of the three cognitive modeling analyses, we used random starting values, and run at least three HMC chains. We collected 4,000, 7,000 and 9,000 samples of each chain after having discarded the first 2,000, 1,000 and 1,000 samples of each chain as burn-in in the case of first, second, and third analysis, respectively. Visual inspection of the chains and confirmation that all parameters had $\hat{R}$ values below 1.05 suggested that the collected samples provided a valid approximation to the joint posterior parameter distribution.

Bayesian hierarchical parameter estimation

Before interpreting the model parameters, we first assessed whether the PVL-Delta model sufficiently accounts for the data of both groups using the post hoc absolute fit method (see Steingroever et al., 2014). The post hoc fit performance of the PVL-Delta model is presented in the third column of Figure 9.1. Comparing the second and third columns of Figure 9.1, it is apparent that the PVL-Delta model nicely captures the qualitative choice pattern in both groups. In particular, as the task proceeds, the model predicts that both groups learn to make more choices from the good decks, and that intuitive decision makers make slightly more choices from the good decks. The PVL-Delta model thus captures key trends in the data, and therefore provides a sufficient account to the data of both groups allowing for meaningful conclusions from the model parameters.

Figure 9.2 shows the posterior distributions of the group-level mean parameters of the PVL-Delta model, separately for the intuitive and the deliberate decisions makers. The posterior distributions show that deliberate decision makers tend to have a higher outcome sensitivity parameter μ_A (i.e., a better correspondence between the objective and the subjective utilities of the decks), but a lower updating parameter μ_a (i.e., less forgetting and weaker recency effects) than intuitive decision makers. In addition, the posterior distributions suggest that the groups differ neither on the loss aversion parameter μ_w nor on the choice consistency parameter μ_c. Note that these conclusions are based only on a visual comparison of the posterior distributions.

Bayes factor model comparison

In this section, we discuss the results of the Bayes factor model comparison. We start by discussing the posterior model probabilities, and then derive Bayes factors using this formula: $\text{BF}_{\Omega,abcd} = \hat{p}(\mathcal{M}_\Omega|D)/\hat{p}(\mathcal{M}_{abcd}|D)$, that is, the ratio of the posterior model probability of model $\mathcal{M}_\Omega$ and $\mathcal{M}_{abcd}$.[4] Tables 9.2 and 9.3 show the posterior model probabilities for eight of the models under the assumption of equal prior model probabilities of all models. The posterior model probabilities of the remaining models are below 0.05 and are not shown. The posterior probability of a specific

[4]The Bayes factors discussed in this article are based on unrounded posterior model probabilities and may therefore slightly differ from Bayes factors calculated using the posterior model probabilities presented in Table 9.2.

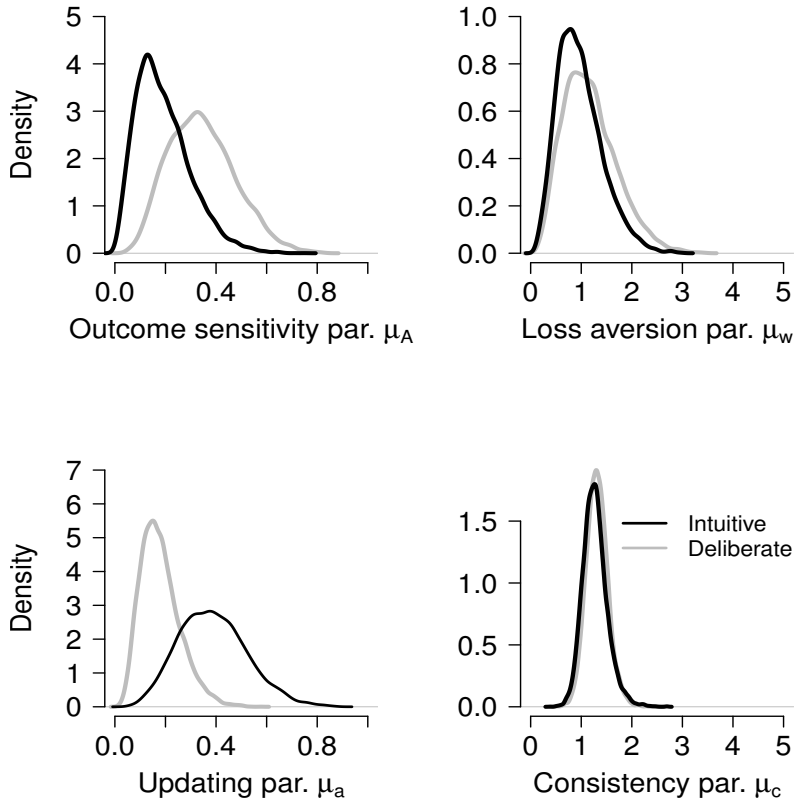

Figure 9.2: Posterior distributions of the group-level parameters of both groups obtained from fitting the PVL-Delta model to the data of each group separately.

model quantifies the evidence that the data provide for that model relative to all other models under consideration (i.e., 15 alternative models). From the tables it is evident that the data provide most evidence for the null model $\mathcal{M}_\Omega$, which assumes no differences between intuitive and deliberate decision makers in the group-level mean parameters. The evidence for the null model is weakest when it is compared to the model that assumes differences between intuitive and deliberate participants in outcome sensitivity parameter (i.e., model $\mathcal{M}_{\mu_A}$; $\text{BF}_{\Omega,\mu_A} = 1.36$) and the model assuming differences in the updating parameter (i.e., model $\mathcal{M}_{\mu_a}$; $\text{BF}_{\Omega,\mu_a} = 1.23$). According to Jeffreys (1961), the evidence for the null model compared to these two models can be characterized as anecdotal. When compared to model $\mathcal{M}_{\mu_w}$ (i.e., the model that assumes differences in the loss aversion parameter), the Bayes factor analysis suggests that there is about three times as much evidence for the null model ($\text{BF}_{\Omega,\mu_w} = 2.84$); according to Jeffreys (1961), this level of evidence is also anecdotal. In addition, the data provide moderate evidence for the null model compared to model $\mathcal{M}_{\mu_c}$ (i.e., the model that assumes differences in the consistency parameter; $\text{BF}_{\Omega,\mu_c} = 6.31$). These findings are consistent with Figure 9.2, where the largest differences in the posterior distributions were on the group-level mean of the outcome sensitivity parameter and the updating parameter; the group-level means for the loss aversion parameter and the consistency parameter had posterior distributions that were highly overlapping.

Table 9.2: *Posterior model probabilities of the null model and models that assume differences in only one group-level mean parameter under the assumption of equal prior model probabilities. The posterior model probabilities of models that are neither shown in this table nor in Table 9.3 are less than .05.*

$\hat{p}(\mathcal{M}_\Omega \mid D)$	$\hat{p}(\mathcal{M}_{\mu_A} \mid D)$	$\hat{p}(\mathcal{M}_{\mu_w} \mid D)$	$\hat{p}(\mathcal{M}_{\mu_a} \mid D)$	$\hat{p}(\mathcal{M}_{\mu_c} \mid D)$
0.20	0.15	0.07	0.16	0.03

Table 9.3: *Posterior model probabilities of models that assume differences in two group-level mean parameters under the assumption of equal prior model probabilities. The posterior model probabilities of models that are neither shown in this table nor in Table 9.2 are less than .05.*

$\hat{p}(\mathcal{M}_{\mu_A\mu_w} \mid D)$	$\hat{p}(\mathcal{M}_{\mu_A\mu_a} \mid D)$	$\hat{p}(\mathcal{M}_{\mu_w\mu_a} \mid D)$
0.05	0.12	0.06

When comparing the null model to models that assume differences in two parameters as in Table 9.3, the null model is generally more strongly supported by the data than when compared to models that assume differences in only one parameter as in Table 9.2. In particular, the data provide anecdotal evidence for the null model compared to the model that assumes differences in both the outcome sensitivity and the updating parameter (i.e., model $\mathcal{M}_{\mu_A\mu_a}$), and moderate evidence for the null model compared to models $\mathcal{M}_{\mu_A\mu_w}$, $\mathcal{M}_{\mu_w\mu_a}$, $\mathcal{M}_{\mu_A\mu_c}$, $\mathcal{M}_{\mu_a\mu_c}$, and $\mathcal{M}_{\mu_A\mu_w\mu_a}$, respectively. For all of the other model comparisons the Bayes factors are greater than 11, suggesting strong evidence for the null model. Thus, our model selection analyses of the data suggest that it is very unlikely that the two groups differ in three or more parameters.

In sum, out of all of the models considered, the null model—that is, the model that assumes no differences in the group-level mean parameters of the intuitive and deliberate decision makers—received most support. Out of all of the models that do assume differences in the group-level mean parameters between the two groups, the model that assumes differences in the outcome sensitivity and the update parameter, respectively, received the most evidence.

Latent-mixture modeling

Figure 9.3 shows the posterior means of the z_i variables for each participant. Since these expectations are naturally interpreted as group membership probabilities, a low posterior mean of z_i suggests that participant i is very likely to belong to the group of deliberate decision makers, whereas a large value suggests that that participant is very likely to belong to the group of intuitive decision makers. According to the group membership established with the decision-style inventory, participants 1–19 were classified as deliberate decision makers (i.e., unfilled bars), whereas participants 20–38 were classified as intuitive decision makers (i.e., grey bars). The horizontal line represents a posterior classification probability of 0.5.

If deliberate versus intuitive decision style has a crucial impact on IGT performance, the latent-mixture model should make inferences consistent with the group membership according to the decision-style inventory. Specifically, for participants 1–19 the posterior mean of the z_i variable should be below the horizontal line, whereas it should be above this line for participants 20–38. However, it is evident in Figure 9.3 that the group membership inferred from the latent-mixture modeling analysis does not coincide with the ground truth distinction between intuitive and deliberate decision makers.

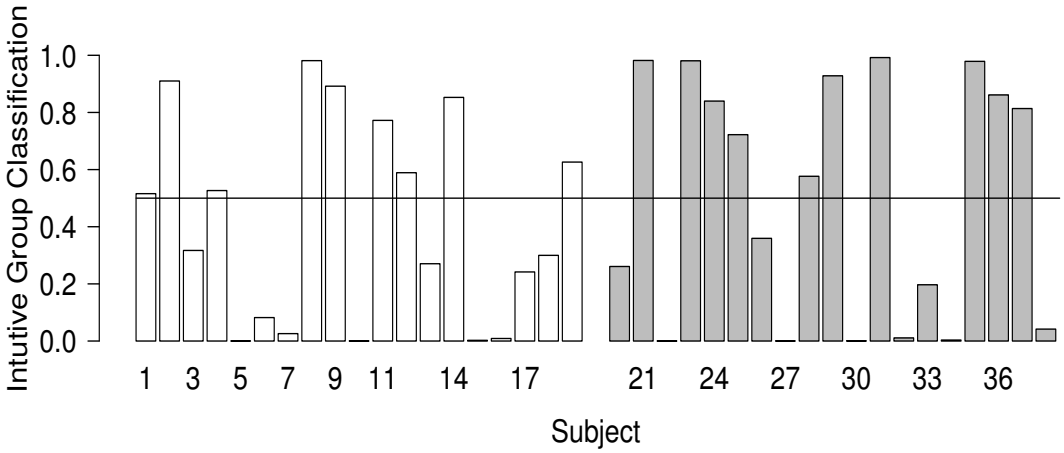

Figure 9.3: Posterior classification as belonging to the group of intuitive decision makers. According to the inventories, participants 1-19 were classified as deliberate decision makers (i.e., white bars), whereas participants 20-38 were classified as intuitive decision makers (i.e., grey bars). The horizontal line represents a posterior classification of .5.

9.4 Discussion

We presented a Bayesian approach for analyzing whether two groups differ in their behavior on the IGT, and for using cognitive models to understand whether their behavior is driven by different psychological processes. For the latter goal, we used three complementary analyses to "triangulate" the research question: Bayesian hierarchical parameter estimation, Bayes factor model comparison, and latent-mixture modeling (see also Lee et al., 2015).

We illustrated this Bayesian approach with a comparison of the choice behavior of intuitive and deliberate decision makers on the IGT. This comparison is interesting because Bechara et al. (1997) proposed that intuitive, affective processes are important for good performance on this task. In addition, intuitive versus deliberate decision style has been found to be linked to behavioral differences in several decision tasks, such as valuation of consumer items and monetary lotteries (Schunk & Betsch, 2006; Betsch & Kunz, 2008). To our knowledge, however, this is the first study that investigates whether decision style also impacts IGT performance.

The application of our Bayesian analysis approach to data from intuitive and deliberate decision makers revealed that, on a behavioral level, intuitive and deliberate decision makers show similar deck preferences on the IGT. Our Bayesian modeling techniques consistently revealed that similar cognitive processes drive performance of intuitive and deliberate decision makers on the IGT. The three different ways of formalizing the basic research question resulted in consistent findings, and, in our view, permit stronger conclusions than could be made based on any one approach individually.

Methodological Contribution

Even though the Bayes factor is "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648), to our knowledge this is the first time that using Bayes factors is proposed to compare IGT performance at both the behavioral level

and using cognitive models. The current standard approach to analyze IGT data is to rely on frequentist methods. In the extensive reviews by Sevy et al. (2007) and Toplak et al. (2010) many non-significant results are reported, not allowing any insightful conclusions, since it can only be concluded that the null hypothesis cannot be rejected. For example, given an alternative hypothesis that states that two groups differ in the number of choices from the good decks as opposed to a null hypothesis that states that there is not such a difference, and a p-value larger than 0.05, one cannot conclude that both groups choose equally often from the good decks. This disadvantage does not hold for the Bayesian approach. The Bayes factor allows an inference about whether the data are informative enough to draw strong conclusions, and, in the case of informative data, further provides an inference about the relative probability of the data under the two competing hypotheses (for more advantages on the Bayesian approach see for example Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008).

Being able to use Bayes factors to compare not only the behavioral performance of two groups (i.e., by means of repeated measures ANOVA), but also to investigate whether two groups differ in PVL-Delta model parameters (i.e., by means of the product space method and latent-mixture modeling) offers a crucial contribution. The current standard in the field of reinforcement-learning models for the IGT is to estimate parameters for each subject separately using maximum likelihood methods, and then use frequentist tests (e.g., independent-samples t-tests, Friedman tests, or Mann-Whitney U tests) to compare the parameter estimates across groups (e.g., Cella et al., 2012; Escartin et al., 2012; Yechiam, Kanz, et al., 2008; Yechiam, Hayden, et al., 2008). However, many studies have shown that maximum likelihood procedures can result in inferior parameter estimates compared to Bayesian hierarchical parameter estimation (Ahn et al., 2011; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, et al., 2010). Thus, the possibility to derive Bayes factors to compare model parameters across groups addresses two major shortcomings of the current standard approach. It allows for a better estimation of the model parameters by avoiding maximum likelihood estimates, and it circumvents shortcomings inherent with classical hypothesis testing.

Theoretical Contribution

Our results offer important insights relating to the discussion of how intuitive and deliberate processes impact decision making. It has been argued that there are robust differences between decision makers in their tendency to rely on the intuitive and the deliberate system (Betsch, 2004), and that these decision styles are linked to behavioral differences in several decision tasks, such as valuation of consumer items and monetary lotteries (Schunk & Betsch, 2006; Betsch & Kunz, 2008). Our results, however, suggest that a person's decision style has no substantial bearing on IGT performance and is thus not a key driver behind the substantial individual differences typically observed in IGT performance.

These results are relevant to previous research in several ways. First, they seem to be inconsistent with the somatic maker hypothesis, according to which a stronger reliance on an intuitive decision mode results in better IGT performance because of the crucial role of the emotional, intuitive system for learning to make good decisions on the IGT (A. Damasio, 1994). Secondly, they also contradict the hypothesis that deliberate decision makers perform better on the IGT because of a strong association between conscious awareness and choosing from the good decks (Maia & McClelland, 2004). Instead, both systems— to the degree that they can be dissociated—seem to equally contribute to performance on the IGT. Finally, our data do not allow one to conclusively answer the question of whether or not deliberate decision makers have a less curved utility function than intuitive decision makers (i.e., a higher outcome sensitivity

parameter; Schunk & Betsch, 2006). On the other hand, note that the weak association between decision style and IGT performance might also be due to the way decision styles are typically assessed. While standard decision-style inventories tap into decision making in the rather abstract and domain-general fashion, there is some indication for considerable domain-specificity of decision style (Pachur & Spaar, 2015). As a consequence, domain-general decision style might only weakly predict decision style in a financial risk task, such as the IGT.

If decision style is only little (if at all) associated with performance on the IGT, what other factors might account for individual variability commonly observed? One possibility is that capacities such as working memory, intelligence, and inhibition play a crucial role. On the other hand, although some studies have indeed found IGT performance to be linked to variables such as working memory, inhibition, intelligence, and personality (e.g., Crone, Vendel, & van der Molen, 2003; Demaree, Burns, & DeDonno, 2010; Franken & Muris, 2005; Suhr & Tsanadis, 2007), such links seem to emerge inconsistently and are, overall, rather weak (e.g., Dunn et al., 2006; Toplak et al., 2010). In light of these results, one cannot rule out that the difficulty in explaining individual differences in IGT performance is also due to characteristics of the task itself. Specifically, Schonberg, Fox, and Poldrack (2011) pointed out that due to its complex nature, tapping into multiple psychological processes (learning, evaluation, and search), the IGT mimics real-world decision making closely. However, this complexity could also have the effect that individual differences unfold in very complex ways on the IGT, making it difficult to identify clear associations between properties of patterns in people's behavior and individual difference measures.

Conclusion

We proposed a set of Bayesian analyses for comparing IGT performance between groups. The application of these techniques to compare decision makers with a deliberate or an intuitive decision style showed not only that both groups of decision makers perform similarly on the IGT, but that their performance is also driven by similar cognitive processes. Our refined analysis approach could easily be adapted to other decision making tasks, and cognitive models of behavior on those tasks. All of the relevant code is available online, and all of the required programs are free to download. Due to the advantages of Bayesian analyses, we encourage using our proposed methodology to investigate group differences in IGT data, or in similar decision-making tasks.

Author Note

The code for all cognitive modeling analyses and the data are available on `www.helensteingroever.com`. The data are also published in Steingroever, Davis, et al. (2015). Additional results, such as tests that confirm the stability of the Bayes factor estimates, and assessment of absolute model account based on the post hoc absolute fit and simulation method for each deck separately, can be requested from the first author. We also repeated the analyses using an alternative classification, which included all participants (distinguishing between intuitive and deliberate decision makers simply based on whether their score on the intuition factor was higher or lower than/equal to the median); results from this analysis can also be requested from the first author.

Acknowledgements

We thank Noemie J. Eichhorn for collecting the data. This research was supported by a Netherlands Organisation for Scientific Research (NWO) grant to HS (404-10-086).