Psychon Bull Rev (2015) 22:391–407
DOI 10.3758/s13423-014-0684-4

THEORETICAL REVIEW

Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice

Benjamin Scheibehenne · Thorsten Pachur

Published online: 19 August 2014

Abstract To be useful, cognitive models with fitted parameters should show generalizability across time and allow accurate predictions of future observations. It has been proposed that hierarchical procedures yield better estimates of model parameters than do nonhierarchical, independent approaches, because the formers' estimates for individuals within a group can mutually inform each other. Here, we examine Bayesian hierarchical approaches to evaluating model generalizability in the context of two prominent models of risky choice—cumulative prospect theory (Tversky & Kahneman, 1992) and the transfer-of-attention-exchange model (Birnbaum & Chavez, 1997). Using empirical data of risky choices collected for each individual at two time points, we compared the use of hierarchical versus independent, nonhierarchical Bayesian estimation techniques to assess two aspects of model generalizability: parameter stability (across time) and predictive accuracy. The relative performance of hierarchical versus independent estimation varied across the different measures of generalizability. The hierarchical approach improved parameter stability (in terms of a lower absolute discrepancy of parameter values across time) and predictive accuracy (in terms of deviance; i.e., likelihood). With respect to test–retest correlations and posterior predictive accuracy, however, the hierarchical approach did not outperform the independent approach. Further analyses suggested that this was due to strong correlations between some parameters within both models. Such intercorrelations make it difficult to identify and interpret single parameters and can induce high degrees of shrinkage in hierarchical models. Similar findings may also occur in the context of other cognitive models of choice.

Keywords Bayesian inference · Parameter estimation · Bayesian modeling · Decision making · Math modeling · Model evaluation

A popular approach in cognitive science to describe and understand behavior is to develop mathematical models with free parameters that can be estimated from empirical data (Busemeyer & Diederich, 2010; Lee & Wagenmakers, 2013; Lewandowsky & Farrell, 2010). The model parameters represent aspects of the assumed underlying psychological processes, which can thus be isolated and quantified. Computational modeling is often used to study individual differences between people and between experimental conditions (e.g., Berkowitsch, Scheibehenne, & Rieskamp, 2014; Pachur & Olsson, 2012) as well as to relate isolated psychological processes to basic cognitive capacities (e.g., working memory; Lewandowsky, 2011; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007) or other individual variables (e.g., aging; Dutilh, Forstmann, Vandekerckhove, & Wagenmakers, 2013).

For illustration, consider cumulative prospect theory (CPT; Tversky & Kahneman, 1992), one of the most prominent models of decision making under risk. According to CPT, responses to risky alternatives (which lead to different outcomes with some probability) are a function of several factors, including a person's sensitivity to outcome and probability information and his or her relative weighting of losses and gains ("loss aversion"). In the model, the degrees of outcome and probability sensitivity and the amount of loss aversion can be quantified by free parameters, and several studies have fitted CPT parameters to investigate how differences in age (Harbaugh, Krause, & Vesterlund, 2002), gender (e.g., Fehr-

Electronic supplementary material The online version of this article (doi:10.3758/s13423-014-0684-4) contains supplementary material, which is available to authorized users.

B. Scheibehenne (✉)
University of Basel, Basel, Switzerland
e-mail: benjamin.scheibehenne@unibas.ch

T. Pachur
Max Planck Institute for Human Development, Berlin, Germany

Duda, De Gennaro, & Schubert, 2006), delinquency (Pachur, Hanoch, & Gummerum, 2010), or affect (Pachur, Hertwig, & Wolkewitz, 2014) influence risky decision making. For such applications of computational modeling, it is often essential to obtain a set of parameter estimates for each individual.

Using and interpreting individually fitted parameters relies on the assumption of model generalizability—that is, the assumption that the fitted model for a person generalizes beyond the circumstances at data collection. For instance, to the extent that the fitted parameters capture stable characteristics of an individual, they should remain relatively invariant across short periods of time or similar contexts, and thus allow prediction of the individual's behavior in those circumstances (e.g., Yechiam & Busemeyer, 2008). However, multiparameter models also run the risk of adjusting, at least in part, to unsystematic variability (i.e., noise). If so, models might overfit the data, meaning that seemingly precise parameter estimates become less meaningful, and consequently "provide less insight and explanation of the cognitive processes they address and are less capable of making accurate predictions when generalized to new or different situations" (Lee & Webb, 2005, p. 606).

The generalizability of a cognitive model for empirical data depends on several factors, including the model's parameterization and the interdependence between its parameters (Li, Lewandowsky, & DeBrunner, 1996), genuine changes in the individual across time, and (unsystematic) error in estimation and measurement. A critical factor influencing this amount of error is the statistical procedure applied to estimate the parameters.

The traditional approach for parameter estimation treats each participant as unique and identifies the best-fitting set of parameter values independently for each individual in a sample. In a second step, these independent estimates can then be aggregated across participants to make inferences about the population from which the individuals were drawn (Gelman & Hill, 2007). As an alternative to this "independence" approach, hierarchical Bayesian procedures have recently grown in popularity (e.g., Gelman & Hill, 2007; Lee & Wagenmakers, 2013; Lee & Webb, 2005; Nilsson, Rieskamp, & Wagenmakers, 2011; Scheibehenne, Rieskamp, & Wagenmakers, 2013). A key characteristic of hierarchical procedures is that they exploit group-level distributions to inform individual-level estimations.

One advantage of hierarchical techniques over the conventional independence approach is their potential to provide parameter estimates that are less prone to measurement error, and thus more stable (Atkinson & Nevill, 1998). This advantage is well justified on theoretical grounds (Gelman & Hill, 2007; Rouder & Lu, 2005), and the hierarchical approach has also proved successful when applied to empirical data (e.g., Rouder, Lu, Morey, Sun, & Speckman, 2008; Scheibehenne & Studer, 2014).

The goal of this article is to rigorously test and compare the potential of hierarchical techniques over an independence approach when assessing parameter stability and the generalizability of decision models across time. This comparison will contribute to a better understanding of what can be gained by using hierarchical techniques and how conclusions regarding the generalizability of a cognitive model that are based on empirical data can be affected by the statistical method applied. One key insight will be that interdependence between model parameters, which is an issue in many cognitive models (e.g., Li et al., 1996; Nosofsky & Zaki, 2002; van Ravenzwaaij, Dutilh, & Wagenmakers, 2011; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010), can lead to substantial differences between the results of hierarchical and independent estimation techniques.

Using data by Glöckner and Pachur (2012), we conducted a series of analyses comparing hierarchical against nonhierarchical (i.e., independent) Bayesian procedures within two prominent models of risky choice: cumulative prospect theory (CPT; Tversky & Kahneman, 1992) and the transfer-of-attention-exchange model (TAX; Birnbaum & Chavez, 1997). We focused on risky choice, because multiparameter models are highly popular in this domain and because it is often assumed that people's responses in risky choice tasks reflect stable individual preferences (Yechiam & Ert, 2011), making it important to know how these potential invariants can best be captured. We compare the results derived from hierarchical and nonhierarchical parameter estimation in terms of two important aspects of model generalizability: (a) the stability of the model parameters—that is, the extent to which estimates remain invariant across time—and (b) the models' predictive accuracy—that is, their ability to predict new data (i.e., data that were not used to inform the parameter estimates).

In the next section, we give an overview of the hierarchical approach to parameter estimation, followed by a formal description of CPT and TAX. We then compare conclusions about the models' generalizability that arise from hierarchical and independent parameter estimationwith respect to parameter stability over time and the accuracy when predicing new data.

Hierarchical versus independent parameter estimation

Hierarchical Bayesian techniques assume that individual parameter values stem from group-level distributions, which are estimated simultaneously with the individual-level parameters. This offers several advantages over the traditional independence approach. In particular, the hierarchical approach naturally lends itself to the hierarchical data structure inherent in many psychological experiments, in which a single individual provides multiple observations and researchers aim to draw conclusions on the aggregate or group level (Gelman, Carlin, Stern, & Rubin, 2004). Unlike independent estimation, Bayesian hierarchical techniques take into account the similarity between individuals and the fact that some individuals might

allow more precise estimates than others; as a consequence, hierarchical techniques can yield more consistent and accurate estimates overall, and thus provide higher statistical power (Rouder & Lu, 2005). This is achieved through partial pooling of the individual estimates on the group level, with the degree of pooling being determined by the data. The purpose of partial pooling is to find an optimal compromise between the extremes of complete pooling and complete independence; the imposed group-level structures simultaneously inform the individual level, such that the individual estimates can borrow strength from the information available about the other individuals in a sample (Gelman et al., 2004).

As was pointed out by Nilsson et al. (2011), this borrowing of strength should increase the reliability of parameter estimates for individual participants, and thus provide more robust results (see also van Ravenzwaaij et al., 2011). In the hierarchical approach, individual parameter estimates that are deemed unlikely given the overall distribution of parameter values (because they are located at the periphery of the distribution) are "corrected" by pulling them closer toward the group mean. This property, sometimes referred to as *shrinkage*, prevents potentially unreliable information from having a disproportionate influence on the group level (Kruschke, 2011). For these reasons, it has been argued that hierarchical methods provide a more thorough and efficient evaluation of models in cognitive science (Shiffrin, Lee, Kim, & Wagenmakers, 2008; van Ravenzwaaij et al., 2011).

For illustration, consider the study by Nilsson et al. (2011) on CPT. On the basis of a model recovery study, they found that "the hierarchical Bayesian method recovered the data-generating parameters somewhat more accurately than [non-hierarchical maximum likelihood estimation]" (p. 89). In addition, the hierarchical approach led to less variable estimates, suggesting greater reliability (and thus generalizability) of CPT. Although these results provide support for the superiority of hierarchical techniques when estimating cognitive models of choice, Nilsson et al.'s (Study 1) simulations were based on the assumption that individuals shared the same set of parameter values. In empirical data, however, parameter estimates often vary greatly between individuals. It is less clear to what degree the shrinkage induced through hierarchical Bayesian modeling will generally yield more reliable estimates in this context. Specifically, it is possible that rather extreme parameter values, which are subject to more shrinkage in the hierarchical approach, represent genuine (and thus stable) characteristics of a person.

To investigate this issue, the data set by Glöckner and Pachur (2012) is particularly suited, because here each participant provided choices between monetary lotteries in two experimental sessions, thus allowing for a genuine test of generalizability (for details, see below). Next we describe CPT and TAX, the two models that we fit to these data.

Cumulative prospect theory

CPT (Tversky & Kahneman 1992) aims to describe how people evaluate risky alternatives that lead to one or several outcomes with some probability. For instance, consider whether you would prefer to play a lottery with a 90 % chance of winning $100 (otherwise nothing) or a lottery with a 10 % chance of winning $1,000 (otherwise nothing). CPT represents a mathematical specification and elaboration of its predecessor, prospect theory (Kahneman & Tversky, 1979), and assumes that the possible consequences of a risky option are perceived as gains or losses relative to a reference point. The overall valuation V of a lottery A with outcomes $x_m > \ldots \geq x_1 > 0 > y_1 > \ldots > y_n$ and corresponding probabilities $p_m \ldots p_1$ and $q_1 \ldots q_n$ is given by:

$$V(A) = \sum_{i=1}^{m} v(x_i)\pi_i^+ + \sum_{j=1}^{n} v\left(y_j\right)\pi_j^-, \tag{1}$$

where v is a *value function* satisfying $v(0) = 0$; π^+ and π^- are the *decision weights* for gains and losses, respectively, which result from a rank-dependent transformation of the outcomes' probabilities. The decision weights are defined as:

$$\begin{aligned}
\pi_m^+ &= w^+(p_m) \\
\pi_n^- &= w^-(q_n) \\
\pi_i^+ &= w^+(p_i + \ldots + p_m) - w^+\left(p_{i+1} + \ldots + p_m\right) \quad \text{for } 1 \leq i < \text{m} \\
\pi_j^- &= w^-\left(q_j + \ldots + q_n\right) - w^-\left(q_{j+1} + \ldots + q_n\right) \quad \text{for } 1 \leq j < \text{n},
\end{aligned} \tag{2}$$

with w^+ and w^- being the probability *weighting functions* for gains and losses, respectively (see below). The weight for each positive outcome is based on the marginal contribution of the outcome's probability to the probability of obtaining a strictly better outcome; the weight for each negative outcome is based on the marginal contribution of the outcome's probability to the probability of obtaining a strictly worse outcome.

Several functional forms of the value and weighting functions have been proposed (see Stott, 2006, for an overview). In our analyses, we use the power value function suggested by Tversky and Kahneman (1992), which is defined as

$$\begin{aligned}
v(x) &= x^{\alpha^+} \\
v(y) &= -\lambda(-y)^{\alpha^-}.
\end{aligned} \tag{3}$$

Values smaller than 1 are usually found for α^+ and α^-, yielding a concave value function for gains and a convex value function for losses. The parameter λ reflects the relative

weighting of losses versus gains and is often found to be larger than 1, indicating loss aversion. In our implementation of CPT, we set $\alpha^+ = \alpha^-$ (and thus had only one α parameter), because Nilsson et al. (2011) found that estimating α^+ and α^- separately can lead to a serious misestimation of λ.

The weighting function has an inverse S-shaped curvature, indicating overweighting of unlikely events (i.e., those with a small probability) and underweighting of likely events (i.e., those with a moderate to high probability). Here, we use the two-parameter weighting function originally proposed by Goldstein and Einhorn (1987), which separates the curvature of the function from its elevation (see Gonzalez & Wu, 1999)[1]:

$$w^+(p) = \frac{\delta^+ p^{\gamma^+}}{\delta^+ p^{\gamma^+} + (1-p)^{\gamma^+}} \qquad w^-(q) = \frac{\delta^- q^{\gamma^-}}{\delta^- q^{\gamma^-} + (1-q)^{\gamma^-}} \tag{4}$$

γ^+ and γ^- (both <1) govern the curvatures of the weighting function in the gain and loss domains, respectively, and indicate the sensitivity to probabilities. The parameters δ^+ and δ^- (both >0) govern the elevations of the weighting function for gains and losses, respectively, and can be interpreted as the attractiveness of gambling. The elevation of the weighting function thus also indicates a person's risk attitude, with higher (lower) values on δ^+ (δ^-) representing higher risk aversion in gains (losses; see Qiu & Steiger, 2011).

In addition to these core components of CPT, a choice rule is required when applying CPT to derive predicted choice probabilities; we used an exponential version of Luce's choice rule (also known as *softmax*; Sutton & Barto, 1998), such that the probability that a lottery A is preferred over a lottery B is defined as:

$$p(A,B) = \frac{e^{\theta \cdot V(A)}}{e^{\theta \cdot V(A)} + e^{\theta \cdot V(B)}}, \tag{5}$$

where θ is a choice sensitivity parameter, indicating how sensitively the predicted choice probability reacts to differences in the valuations of lotteries A and B. A higher θ indicates more deterministic behavior; with $\theta = 0$, choices are random. To summarize, CPT as implemented here has seven free parameters: outcome sensitivity (α), loss aversion (λ), separate probability sensitivities for gains and losses (γ^+, γ^-), separate elevations for gains and losses (δ^+, δ^-), and choice sensitivity (θ).

[1] Note that Nilsson et al. (2011) used Tversky and Kahneman's (1992) one-parameter probability-weighting function, which does not disentangle elevation and curvature.

Transfer-of-attention-exchange model

TAX (Birnbaum & Chavez, 1997) provides an alternative model to CPT that can account for some empirical violations of CPT (for a summary, see Birnbaum, 2008). According to TAX, the valuation of a lottery is a weighted average of the utilities of the outcomes; the weight that each outcome receives depends on its rank among all possible outcomes (the n outcomes being ordered such that $x_1 < x_2 < x_3 < \ldots x_n$) and its probability. To account for the typically observed risk aversion (risk seeking) in gains (losses), the model assumes that attention (i.e., weight) is "transferred" from better (worse) to worse (better) outcomes. Specifically, the valuation V of lottery A is calculated as

$$V(A) = \frac{\sum_{i=1}^{n}\left[t(p_i) + \frac{\delta}{n+1}\sum_{j=1}^{i=1} t\left(p_j\right) - \frac{\delta}{n+1}\sum_{j=1}^{n} t(p_i)\right]u(x_i)}{\sum_{i=1}^{n} t(p_i)}, \tag{6}$$

where δ is a free parameter governing the attention shift from higher to lower outcomes (or vice versa); with $0 < \delta < 1$, attention is shifted from higher (lower) to lower (higher) outcomes in gains (losses); with $0 > \delta > -1$, the opposite occurs. In the equation, $u(x)$ is the utility function

$$u(x) = x^{\beta} \quad if \; x > 0 \qquad u(x) = -(-y)^{\beta} \quad if \; x \le 0 \tag{7}$$

transforming objective outcomes into subjective utilities. The free parameter β indicates the curvature of the value function and reflects the decision maker's sensitivity to outcome information (with lower values of β indicating lower sensitivity). Finally, $t(p)$ is the probability-weighting function, transforming objective into subjective probabilities, and equals

$$t(p) = p^{\gamma}. \tag{8}$$

where γ is a free parameter reflecting the decision maker's sensitivity to probability information (with lower values of γ indicating lower sensitivity). As for CPT, the predicted probability that lottery A is preferred over lottery B is derived using the softmax rule (Eq. 5). To summarize, TAX as implemented here has four free parameters: attention shift (δ), outcome sensitivity (β), probability sensitivity (γ), and choice sensitivity (θ).

Overview of the analyses

We applied both CPT and TAX to model the data reported in Glöckner and Pachur (2012). These data offer an expedient

context for assessing model generalizability, because participants gave responses to a large and varied set of lottery problems, allowing parameter estimation for each participant. Moreover, each participant provided responses at two experimental sessions, allowing us to examine parameter stability and predictive accuracy. At each session (t1 and t2, respectively, which were separated by one week), each of 64 participants (39 female, 25 male; mean age 24.7 years) were presented with 138 two-outcome monetary lotteries covering pure-gain, pure-loss, and mixed outcomes. All of the lotteries were drawn from three sets of lottery problems used in previously published studies. Thirty-eight of the problems were shown at both sessions; the other 100 were drawn in equal proportions from the three sets. The outcomes of the lotteries ranged from –€1,000 to €1,200. At the end of each session, one of the chosen lotteries was picked randomly, played out, and the participant received payment proportional to the outcome (according to a specific exchange rate).

Using a maximum likelihood estimation procedure and the independence approach, Glöckner and Pachur (2012) found that CPT's parameters were relatively stable (as measured by test–retest correlations) across the two experimental sessions; moreover, CPT predicted the individual choices across time rather well (as measured by the percentage of correct predictions). Extending these analyses, we used both hierarchical and independent Bayesian techniques to estimate the parameters of CPT and TAX to compare the conclusions on two key aspects of model generalizability. First, we compared both estimation approaches on two prominent measures of parameter stability—namely, test–retest correlations and coefficients of variation. Second, we tested the extent to which the estimation approaches yielded parameter sets that differed in their ability to predict new data (including both posterior predictive accuracy and accuracy in out-of-sample predictions). To conduct these comparisons, we implemented both hierarchical and independent Bayesian versions of CPT and TAX, which are outlined next.

Model specification and Bayesian parameter estimation using BUGS

The free parameters of CPT and TAX were estimated using Bayesian versions of each model implemented in the BUGS programming language (Lunn, Spiegelhalter, Thomas, & Best, 2009).[2] Bayesian procedures require a detailed specification of a model, including its respective likelihood function and the prior probability distributions of all estimated parameters. For the independent versions of CPT and TAX, we specified the models on the basis of Eqs. 1–5 and Eqs. 5–8, respectively. The priors for the free parameters were set to uniform probability distributions spanning a reasonable range that excluded theoretically implausible values and allowed ample space to include parameter values obtained in previous research. For CPT, the priors ranged from 0 to 5 for θ, λ, δ^{+}, and δ^{-}, and from 0 to 1 for α, γ^{+}, and γ^{-} (see Glöckner & Pachur, 2012; Rieskamp, 2008). For TAX, the priors ranged from –2 to 2 for δ and from 0 to 5 for β, γ, and θ (Birnbaum, personal communication, 19th of July 2012).

In the hierarchical versions of CPT and TAX, we used the same functions as in the independent versions, but partially pooled the individual parameters through group-level distributions. Priors in the hierarchical version were set such that the range and the (uniform) shape of the prior distributions on the individual level matched those in the independent version. Toward that goal, normally distributed group-level parameters were linked to the individual level through probit transformations to achieve a uniform distribution from 0 to 1 (Rouder & Lu, 2005). To extend the range of these distributions on the individual level from –2 to 2 for δ and from 0 to 5 for β, γ, and θ, we interposed an additional linear linkage function. All hierarchical group-level means were assumed to be normally distributed with a mean of 0 and a variance of 1 (yielding uniform distributions on the individual level). The priors on the group-level standard deviations were uniformly distributed, ranging from 0 to 10 (thus avoiding extreme bimodal distributions on the individual level).

For both the individual and the hierarchical models, we estimated the joint posterior parameter distributions using Monte Carlo Markov chain methods implemented in JAGS, a sampler that utilizes a version of the BUGS programming language (version 3.3.0; Plummer, 2003) that was called from the R statistics software (version 2.14.0; R Development Core Team, 2012). A total of 40,000 representative samples were drawn from the posterior distributions after a "burn-in" period of 1,000 samples. The sampling procedures were efficient, as indicated by low autocorrelations of the sample chains, the Gelman–Rubin statistics, and visual inspections of the chain plots.

Parameter estimates

Both the independently and the hierarchically estimated parameter values were well within the bounds of the assumed prior range. The posterior distributions of the independent estimates were more dispersed than the hierarchical ones, which is not surprising, given that the independent estimates

[2] The BUGS programming code for each model implementation is available in the online supplementary materials.

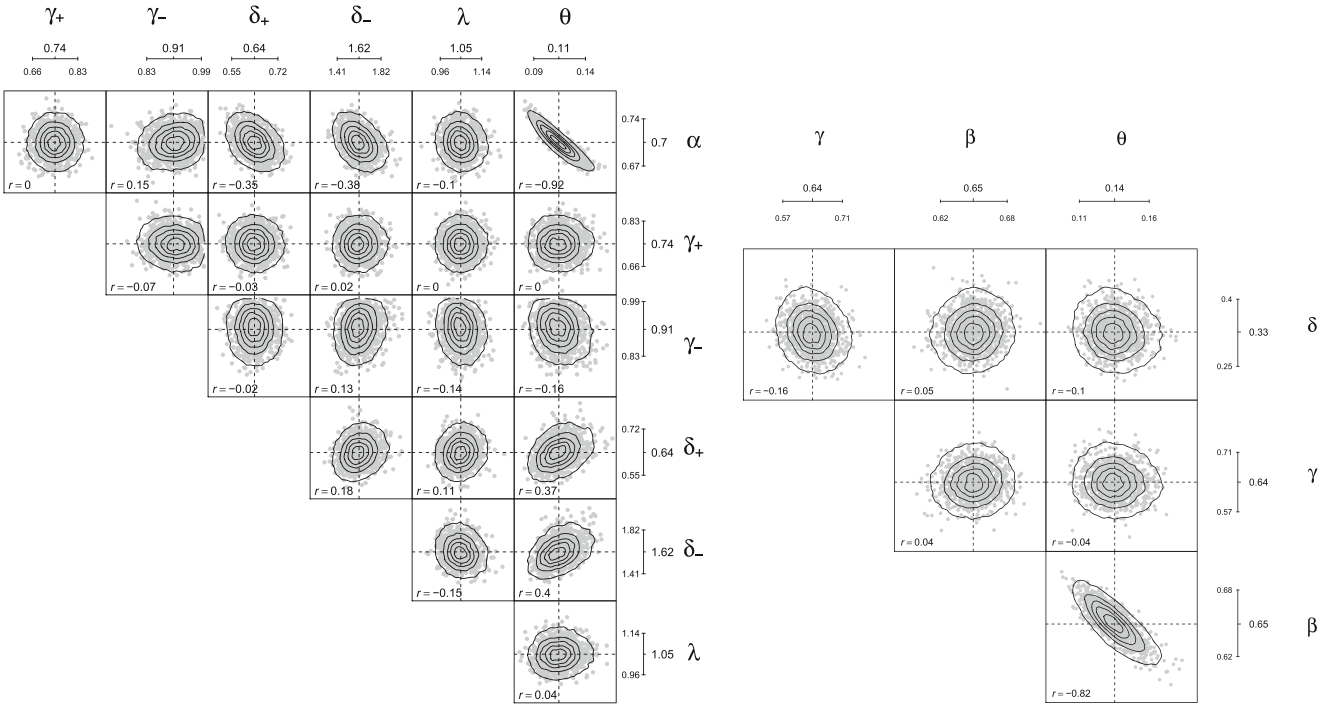

Fig. 1 Hierarchically estimated posterior parameter distributions for cumulative prospect theory (CPT, left plots) and the transfer-of-attention-exchange model (TAX, right plots) on the group level. Each square displays the joint posterior distributions for any pair of parameters along with the respective product-moment correlation coefficients. Grey dots are 1,000 random samples from the posterior distribution. The error bars at the top and right margins indicate the group-level mean and 95 % highest posterior density interval (HDI_{95}) for each parameter

were not partially pooled through hierarchical group-level distributions.

Figure 1 shows the hierarchically estimated group-level means of the parameters for both CPT and TAX in the first experimental session (i.e., t1) of Glöckner and Pachur (2012). The independently estimated mean and median parameter values are shown in Tables 1 and 2. Overall, the results are comparable to previously obtained parameter estimates (see Fox & Poldrack, 2008, and Birnbaum, 2008, for CPT and TAX, respectively), and the independently estimated parameters also approximate those from Glöckner and Pachur's analysis based on maximum likelihood estimation. Both CPT and TAX show the typical pattern of reduced outcome sensitivity (with α and β both being <1) and probability sensitivity (with γ^+/γ^- <1 and γ <1).

Comparisons of the hierarchical and independent estimates reveal some notable differences. First, for CPT, the hierarchical approach leads to lower values of the loss aversion parameter λ. Second, the hierarchical approach suggests a different amount of risk aversion: The elevation of CPT's weighting function (as indexed by δ^+ and δ^-) is lower than that from the independent approach, indicating lower risk aversion in gains and higher risk aversion in losses. The same holds for TAX, with a lower attention shift parameter δ than in the hierarchical estimates. Third, for both CPT and TAX, the hierarchical estimates suggest a lower choice sensitivity, as indicated by lower estimates of the θ parameter.[3] We will come back to this difference in θs when assessing the predictive accuracy of the models.

Figure 1 further indicates that, for both models, some parameters were substantially correlated on the group level. This makes it difficult to interpret parameter values independently (see Li et al., 1996). Correlations were particularly high between the group-level means of the choice sensitivity parameter θ and the outcome sensitivity parameters, for both CPT (α; r = –.92) and TAX (β; r = –.82). On theoretical grounds, these interdependencies seem plausible, because the sensitivity parameter scales the subjective valuations V of the available options (see Eqs. 3 and 7), whereas θ determines how sensitive decision makers are to differences between these valuations (in the softmax choice rule in Eq. 5, V is multiplied by θ). Thus, if subjective valuations increase, a smaller θ value would be required to maintain the predicted choice probabilities. As will be elaborated later in this article,

[3] Because Nilsson et al. (2011) employed a different weighting function in their comparison of independent and hierarchical parameter estimations for CPT, our results are not directly comparable with theirs. Nevertheless, note that Nilsson et al. also found the hierarchical approach to yield a lower choice sensitivity. Interestingly, they obtained a pattern of results opposite to ours with regard to loss aversion, with the hierarchical approach yielding a higher λ.

Table 1 Independent parameter estimates for cumulative prospect theory, shown separately for the two experimental sessions (t1 and t2) in Glöckner and Pachur (2012)

		α	λ	γ^+	γ^-	δ^+	δ^-	θ
t1	*M*	.53	1.30	.68	.67	0.95	2.38	0.72
	Md	.52	1.12	.69	.70	0.84	2.46	0.46
t2	*M*	.55	1.21	.63	.67	0.87	2.38	0.70
	Md	.57	1.20	.64	.75	0.69	2.22	0.44

Table 2 Independent parameter estimates for the transfer-of-attention-exchange model, shown separately for the two experimental sessions (t1 and t2) in Glöckner and Pachur (2012)

		β	δ	γ	θ
t1	*M*	0.56	0.42	0.71	0.58
	Md	0.57	0.35	0.57	0.33
t2	*M*	0.55	0.48	0.74	0.69
	Md	0.55	0.43	0.53	0.35

these parameter interdependencies are critical for the comparison of hierarchical and independent estimation approaches.

Do hierarchical estimates lead to higher parameter stability?

To the extent that the parameters of a cognitive model capture stable characteristics of individual participants, the individual participants' parameter values should be invariant across time—at least for relatively short time intervals and under comparable measurement conditions (Bland & Altman, 1986). To test whether the Bayesian hierarchical approach gives rise to more stable parameter estimates than does the independent approach, we calculated two measures of parameter stability: test–retest correlations and the coefficients of variation.

Test–retest correlation

Perhaps the most common way to quantify stability (or reliability) is to calculate correlations between two measurement points in time (see Hopkins, 2000). To assess this test–retest reliability for the parameters of CPT and TAX, we computed Pearson's product-moment correlations across t1 and t2 for the posterior means of all individual parameters. To compare the two estimation procedures, we conducted the reliability analysis once for the hierarchically estimated parameters and once for the independently estimated parameters. The correlations were calculated using Bayesian techniques implemented in BUGS (which avoid many problems inherent in traditional frequentist procedures that rely on null-hypothesis significance testing; Kruschke, 2011).[4]

Figure 2 displays the test–retest correlations, *r*, for the parameters of CPT (left plot) and TAX (right plot). As can be seen, for most parameters the highest posterior density intervals (HDI_{95}) overlap, indicating no credible difference between the correlations of the hierarchically and the independently estimated parameters; this holds for both CPT and TAX. If anything, the correlations are slightly lower for the hierarchical estimates, and in two cases (the θ parameters of CPT and TAX), they are even credibly lower. A similar picture emerged when we used Spearman's rank correlations as a measure of reliability (not shown). Thus, with respect to test–retest correlations, we found no consistent advantage of hierarchical over independent techniques.

Coefficient of variation

Although correlations are a popular measure of association, they provide only a relative measure of reliability; they do not capture the extent to which two variables agree in absolute terms (Atkinson & Nevill, 1998). For example, two measures can be perfectly correlated even when the absolute differences between them are large. Furthermore, correlational measures can be difficult to interpret because they are sensitive to heterogeneity among participants and to the range of the values (Hopkins, 2000). One way to quantify the reliability of a parameter in absolute terms is to calculate the standard deviation (across participants) of the differences between each participant's parameter values at t1 and t2. To obtain an interpretable standardized measure, this standard deviation can be divided by the average parameter value (across participants). Expressed as a percentage, this index is referred to as the *coefficient of variation* (CV; Hendricks & Robey, 1936). Thus, the CV can be seen as a measure of the similarity of two measurements (see also Hopkins, 2000). Applied to the present case, a CV of 30 % would indicate that, on average across all participants, the two parameter values differed by 30 % of the mean. In contrast to alternative distance measures, such as root-mean-squared deviation, the CV has the advantage of being scale-independent (i.e., its magnitude does not depend on the absolute parameter values). This makes it possible to compare the stability of parameters that differ in magnitude. To compare the two estimation procedures on the basis of CV, we used Bayesian techniques implemented in BUGS and calculated the CVs separately for the hierarchically estimated and the independently estimated parameters.

As is shown in Fig. 3, for both CPT and TAX the hierarchically estimated parameters show credibly lower CVs (indicating higher reliability) than the independently estimated parameters. Thus, hierarchical techniques seem to yield

[4] See the online supplementary materials for the BUGS programming code.

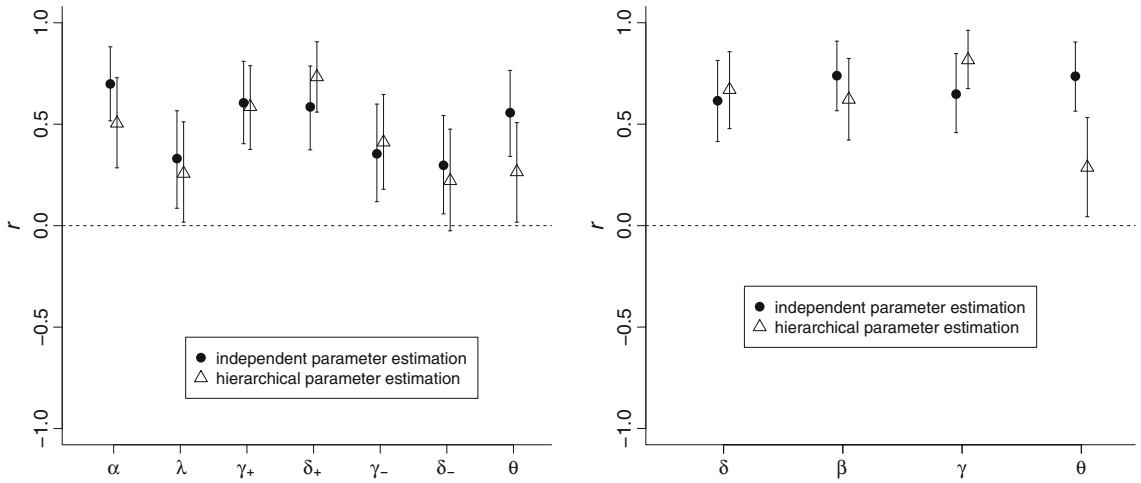

Fig. 2 Estimated test–retest correlations for the individual parameters of CPT (left plot) and TAX (right plot). Filled circles and open triangles indicate the mean posterior estimates for independent and hierarchical parameter estimates, respectively. Error bars are HDI_{95}s

parameter estimates that are more reliable in absolute terms, such that estimates taken at two different points in time will differ less relative to the parameters' mean.

Why did the hierarchical approach impact primarily on absolute measures of parameter stability?

The results indicate that hierarchical techniques lead to more similar parameter estimates in absolute terms (as measured by the CV), but not consistently to higher test–retest correlations. The latter result may seem surprising, because the supposed advantages of hierarchical techniques, which borrow strength from distributional information on the group level, should generally yield more reliable parameter estimates on the individual level.

Lower correlation due to shrinkage

To gain a better understanding of this result, we examined the distributions of the individual parameter estimates more closely. Figure 4 displays the posterior means of the β parameter in the TAX model for a subset of 20 representative participants.[5] Here, the independently estimated parameter values at t1 and t2 are plotted along the upper and lower rows; the hierarchically estimated parameters at t1 are displayed along the middle row. The estimates for each participant are connected by a line. As can be seen, the partial pooling of the hierarchical approach led to a clearly lower dispersion of the estimates than for the individually estimated parameters (the same holds for the hierarchical estimates at t2, which are not shown). Figure 4 further shows that shrinkage is particularly strong for extreme parameter estimates—that is, for those that are far away from the group-level mean. The reason is that these estimates appear rather unlikely, given the group-level distribution, and are thus implicitly treated as extreme values in the hierarchical model. Importantly, Fig. 4 also shows a fairly close correspondence between the independent estimates at t1 and t2, even for participants with rather extreme parameter values. That is, individuals who score high on the β parameter at t1 also tend to score high at t2; the same applies for small β values. This indicates that, for the present data, extreme estimates can reflect meaningful and reliable characteristics of the individuals.

To illustrate the consequences of the shrinkage induced by the hierarchical approach for the test–retest correlations, Fig. 5 displays a scatterplot for the θ parameter in the TAX model separately for the independent and the hierarchical estimates (this example is instructive because, as is shown in Fig. 2, the difference between the correlations for the individual and the hierarchical estimates was particularly large here). As can be seen, the high correlation for the independent estimates (upper plot) is partly due to some individuals who have high values on the θ parameter at both t1 and t2. Inspection of the distribution of the hierarchically estimated parameters (lower plot) shows, due to shrinkage, a considerably narrower value range than in the case of the independent estimates (note that the axis scales in Fig. 5 were adjusted to facilitate display of the data). The result is a lower (linear) correlation between the two measurement points than in the case of the independently estimated parameter values (which left the extreme parameter values intact).[6]

[5] See the online supplemental materials for similar plots of the other parameters.

[6] Note, however, that this would not necessarily be the case. It is possible to conceive of situations in which shrinkage reduces the variance but retains the (linear) relationship between the individual parameters; in such cases, the test–retest correlations would not be lower for hierarchically estimated parameters, as they indeed are not for most of the parameters in Fig. 2.

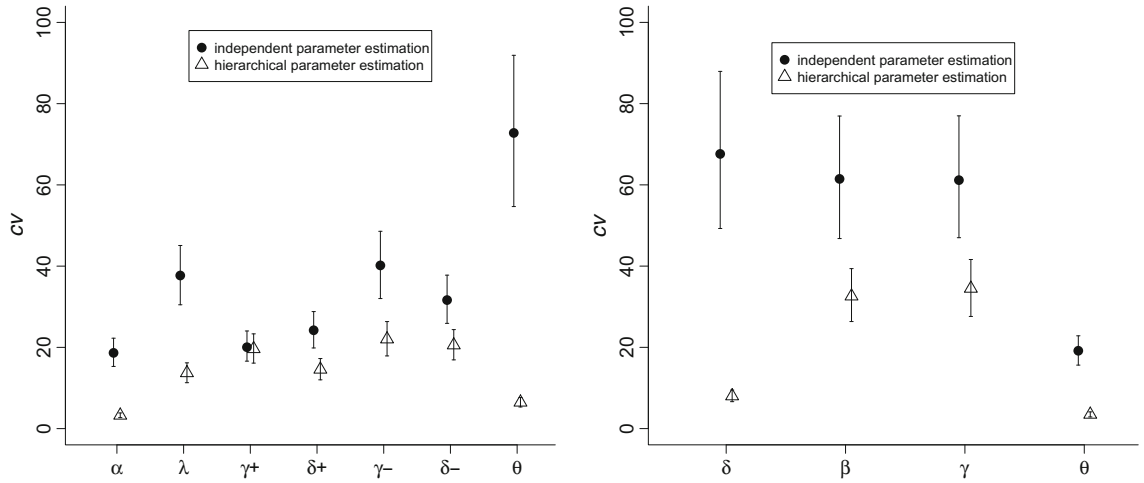

Fig. 3 Estimated coefficients of variation (CVs) for the individual parameters of CPT (left plot) and TAX (right plot). Filled circles and open triangles indicate the mean posterior estimates for independent estimates and hierarchical estimates, respectively. Error bars are HDI_{95}s

Figures 4 and 5 also hint at why the hierarchical approach nevertheless leads to an improvement of absolute reliability, such as the CV. As can be seen from the figures, the variance of the independently estimated parameters is considerably higher than that of the hierarchically estimated parameters (which scatter closely around the group-level mean). The higher variance in the independent case is further increased by a few extreme estimates. These extreme cases presumably boost the estimate of the variance more than they boost the mean, and consequently lead to a higher CV.

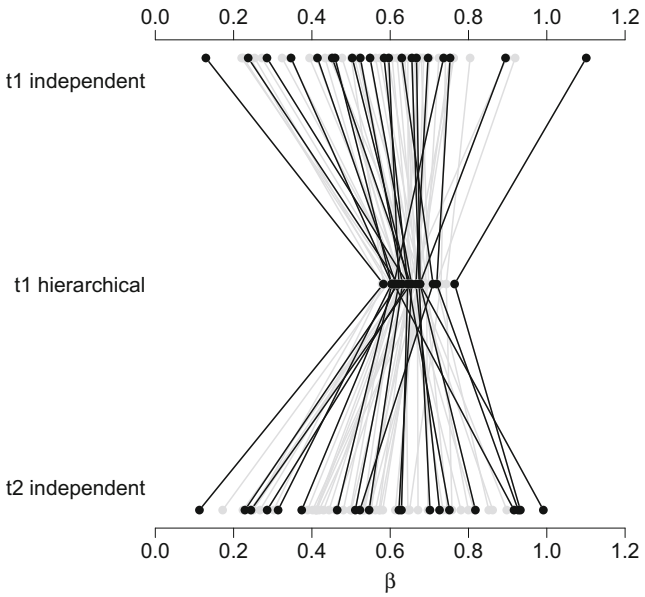

Fig. 4 Mean posterior estimates of the β parameter of TAX, shown separately for each individual at t1 and t2 (upper and lower row), and the hierarchically estimated parameters at t1 (middle row). For illustrative purposes, the dark points highlight a subset of 20 participants across the data range. The gray dots represent the remaining participants within the sample

Higher shrinkage for interdependent model parameters

Although in our analysis the shrinkage imposed through hierarchical modeling led to a lower test–retest correlation overall, the same did not apply to all parameters. As can be seen in Fig. 2, for some parameters the correlations were slightly higher for the hierarchical than for the independent estimates (e.g., δ^+ in CPT and γ in TAX). What might determine whether or not the hierarchical approach decreases a test–retest correlation? One key factor may be the degree of interdependence between parameters: Those parameters for which the discrepancy between the hierarchical and independent estimates was most pronounced were also those showing the strongest intercorrelation (Fig. 1)—namely, the choice sensitivity parameter (θ) and the outcome sensitivity parameter (i.e., α and β for CPT and TAX, respectively). High correlations between parameters imply that different combinations of these parameters are about equally probable and that the marginal posterior distribution of each single parameter is rather dispersed. As a consequence, when estimated hierarchically, these parameters may be more susceptible to shrinkage toward the group-level mean. In support of this hypothesis, when we reestimated a reduced version of TAX in which the values of the individual θ parameters were fixed to the (previously estimated) posterior means, the shrinkage of the β parameter was far less pronounced.

In summary, we found that hierarchical Bayesian techniques do not necessarily yield higher test–retest correlations than do individually estimated parameters. This seems to hold particularly for parameters that show strong interdependence, highlighting the important role of the formal architecture of a model when pursuing a hierarchical estimation approach. As we will outline in the next section, the effect of shrinkage in the hierarchical approach also has consequences for a model's predictive accuracy.

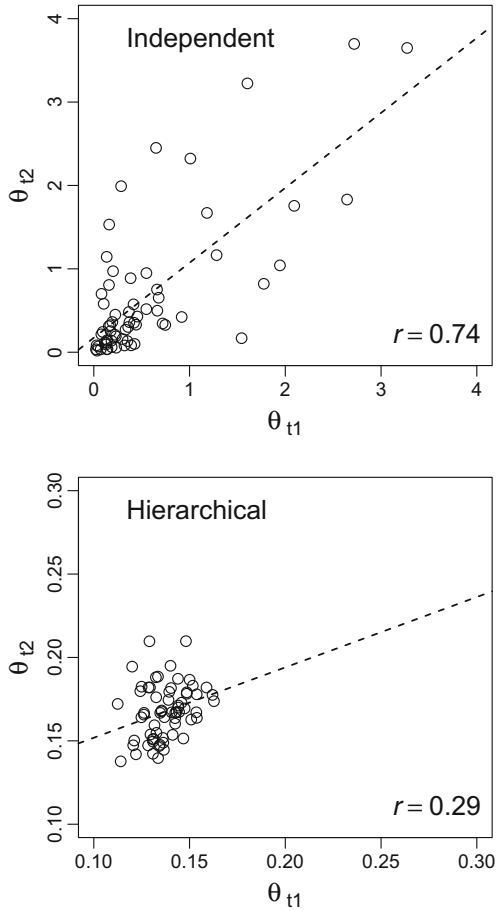

Fig. 5 Scatterplot of the mean posterior estimates for the θ parameter in TAX at t1 and t2. Each point represents one participant. The upper panel shows the parameter values obtained through independent estimation; the lower panel shows the parameter values obtained through hierarchical techniques. Note that the value ranges on the axes are much smaller in the lower panel

Do hierarchically estimated models make better predictions?

Although the stability of parameter estimates across time is an important aspect of a model (Glöckner & Pachur, 2012), a further and perhaps more direct test of model generalizability is the degree to which it makes correct predictions. The abilities of CPT and TAX to predict new data can be tested by comparing participants' choices at t2 against predictions based on the parameters obtained at t1. To the extent that hierarchical techniques reduce error variance in the parameter estimates, using these parameters should lead to more accurate predictions than using independently estimated parameters. In the following, we compare hierarchical and independent estimates in terms of two aspects of predictive accuracy: (a) predictions across time for the same individual and (b) out-of-sample predictions for "new" individuals based on the mean parameters on the group level. As measures of predictive accuracy, we estimated the posterior predictive accuracy as well as the deviance.

Posterior predictive accuracy

Posterior predictive checks were applied using a two-step approach (Gelman et al., 2004, pp. 157ff). First, we generated model predictions for each individual by randomly sampling 4,000 parameter values from the respective joint posterior distributions obtained at t1. In a second step, we estimated the mean probability with which the model predicted the actually observed choices at t2 for the same individual across all parameter samples. This fit measure is sometimes referred to as the *posterior predictive probability* (p; Gelman et al., 2004). Here, $p = 1$ indicates perfect predictive accuracy, whereas $p = .5$ indicates accuracy no better than chance.

Figure 6 plots for each participant the posterior predictive probability calculated from independently estimated parameters (x-axis) against those estimated hierarchically (y-axis), shown separately for CPT (left plot) and TAX (right plot). As can be seen, the posterior predictive probability is higher than .5 for all participants, indicating that the predictive accuracy of both models was above chance level. Furthermore, the points scatter equally around both sides of the diagonal, indicating that the proportions of correct predictions for the hierarchical and independent approaches were comparable. Across participants, the mean differences between the approaches (calculated as $p_{\text{independent}}$ minus $p_{\text{hierarchical}}$) were –.0028 (SD = .02) for CPT and .0037 (SD = .02) for TAX. Both differences are rather small and not significantly different from zero [CPT: $t(63) = 1.13$, $p = .262$, $BF_{01} = 5.4$ (i.e., the Bayes factor indicates that the data was 5.4 times more likely under the null hypothesis); TAX: $t(63) = 1.49$, $p = .141$, $BF_{01} = 3.4$]. Taken together, hierarchical and independent estimation techniques yielded comparable accuracy when we used the estimated models to predict new data for individual decision makers on the basis of their choices at a previous point in time on similar lottery problems.

Out-of-sample predictions from group-level estimates

As a further measure of predictive accuracy, we examined the models' abilities to predict the choices of an individual participant on the basis of observed data from other participants in the sample. Thus, in contrast to the previous analysis, in which we tested the predictive accuracy for choices observed at t2 on the basis of parameters estimated at t1 for the same individual, the goal was now to predict choices of a randomly drawn participant at t2 on the basis of the group-level means estimated at t1. Here, the question was whether hierarchically estimated group-level means or simply the arithmetic mean of the independently estimated parameters lead to better out-of-sample predictions.

One possible advantage of using hierarchical estimation for out-of-sample predictions is that it may yield more reliable parameter estimates on the group level. Unlike in the context

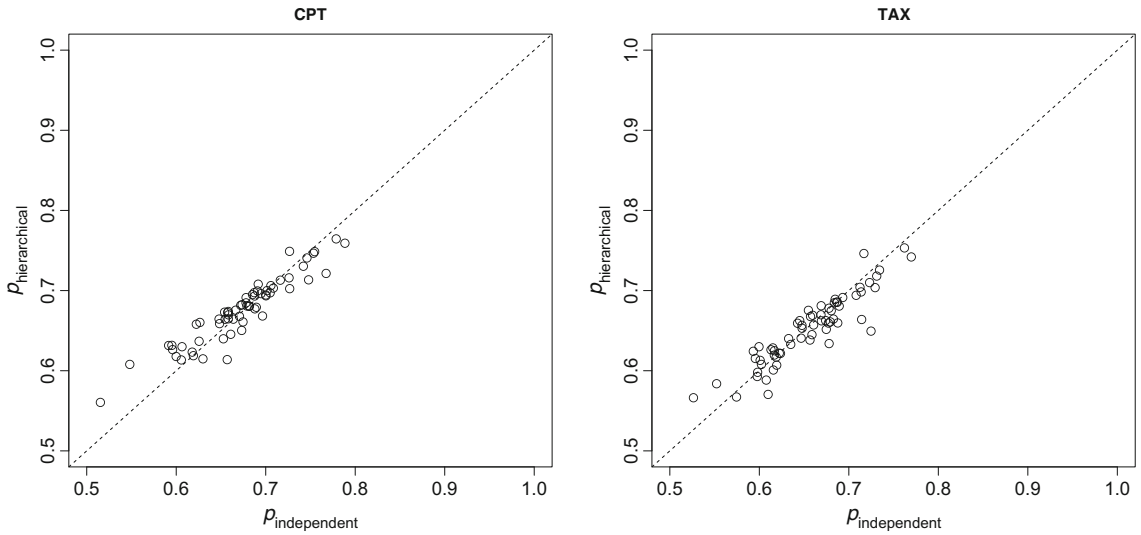

Fig. 6 Mean posterior predictive probabilities of individual parameter distributions that were estimated independently (*x*-axis) or hierarchically (*y*-axis). The left plot displays results for CPT, and the right plot displays results for TAX. Dots below the diagonal indicate higher predictive accuracy for independent estimates

of the arithmetic mean—where each individual value contributes equally—parameters in (Bayesian) hierarchical models that are estimated with less precision (i.e., that have higher variance) and parameters that are farther away from the group-level mean receive less weight. Because of this, hierarchical means can be considered more representative of the group as a whole, and may thus be superior when predicting the behavior of new, previously unobserved group members.

To test this prediction, we first tested the hierarchical approach by calculating the posterior predictive probability for each individual participant at t2 on the basis of the posterior group-level parameters estimated at t1 for both CPT and TAX. For the independent approach, the procedure was similar, but the parameter values were obtained from the arithmetic means of the respective individual parameters, calculated across participants at t1. For each individual prediction, the participant whose choices were predicted at t2 was excluded from the averaging of the independently estimated parameters at t1.[7]

For CPT and TAX, the mean posterior predictive probabilities for the hierarchical predictions were 4.1 (*SD* = 2.6) and 4.5 percentage points (*SD* = 1.9), respectively, lower than those for the independent predictions; both differences are statistically significant, as indicated by conventional *t* tests [for CPT, $t(63) = 12.7$, $p < .001$, $BF_{10} > 10{,}000$ (i.e., the Bayes factor indicates that the data are over 10,000 times more likely under Hypothesis 1); for TAX, $t(63) = 18.4$, $p < .001$, $BF_{10} > 10{,}000$]. A similar picture emerged when we made single-point predictions based on the means of the respective group-level parameters rather than drawing many representative samples from the posterior group-level distributions.

Predictive accuracy measured by deviance

An alternative criterion for assessing predictive accuracy is in terms of likelihood, defined as the product of the probabilities of the actually observed data at t2 for a given parameterization of the model based on t1. Because multiplying probabilities often yields very small values, a common approach is to report likelihood in terms of *deviance*, defined as –2 times the sum of the logarithms of each likelihood. In general, the larger the discrepancies between the predicted and the observed choices, the higher the deviance will be (indicating a lower likelihood). To compare the abilities of hierarchical and independent estimation to predict new group members in terms of deviance, we took an approach similar to the one in the analysis above: We first generated model predictions for the observed individual choices at t2 based on the parameter estimates at t1; we then calculated the deviance by comparing the observed choices against the probabilistic predictions of each model.

When we predicted individuals' choices at t2 on the basis of their parameter values at t1, the independent and hierarchical approaches yielded comparable results for TAX. Here, the mean deviance based on the independent parameter estimates was 149 (*SD* = 18.3), as compared to 146 (*SD* = 16.9) based on the hierarchical estimates, $t(63) = 2.17$, $p = .034$, $BF_{10} = 0.9$.[8] For CPT, the mean deviance of the independent predictions (149, *SD* = 29.2) was higher than that of the hierarchical

[7] For pragmatic reasons, in the hierarchical case all participants, including those predicted at t2 at any one time, were included in the parameter estimation. This may have yielded a small advantage for the hierarchical approach over the independent approach.

[8] Bayes factor estimates were calculated from conventional *t*-test outputs on the basis of the template by Rouder, Speckman, Sun, Morey, and Iverson (2009), assuming the Jeffrey–Zellner–Siow prior and $r = 1$.

predictions (141, *SD* = 23.2), indicating higher predictive accuracy for the latter, $t(63) = 5.43$, $p < .001$, $BF_{10} > 10{,}000$. For both TAX and CPT, the mean deviance was clearly lower than 191, the deviance expected under random choice.

For out-of-sample predictions (i.e., predicting a participant's choices at t2 on the basis of the mean posterior parameters across the other participants at t1), the advantage of the hierarchical approach in terms of deviance was even more pronounced. As is shown by Fig. 7, which plots the deviance of the predictions based on the mean of the independent parameter estimates against the deviance of the hierarchical mean parameter estimates, the deviance is generally smaller (indicating better fit) for the hierarchical parameter estimates. (Note that, in contrast to Fig. 6, a larger value here means poorer performance.) For CPT, the mean deviance based on the hierarchical parameters is 187 (*SD* = 57.9), as compared to 147 (*SD* = 25.3) based on the independent parameters, yielding a difference of 41. For TAX, the means are 151 (*SD* = 20.2) and 209 (*SD* = 51), yielding a difference of 58. Both differences are statistically significant, as indicated by conventional *t* tests [for CPT, $t(63) = 8.9$, $p < .001$, $BF_{10} > 10{,}000$; for TAX, $t(63) = 13.9$, $p < .001$, $BF_{10} > 10{,}000$]. For quite a few individuals, the deviance for the independent estimates was even higher (i.e., worse) than would be expected under chance, suggesting that some choices were particularly poorly predicted. Presumably this occurred because, although the independent models made correct predictions most of the time, the predictive accuracy for a few observed choices was extremely low, hence escalating deviance. The hierarchical parameters, by contrast, produced less extreme choice probabilities (because they represent an average across all individuals), thus avoiding extreme prediction errors.

Why do the results for posterior predictive accuracy and deviance diverge?

At first glance, the lower posterior predictive accuracy of the hierarchical models may seem surprising, given that the hierarchical approach outperformed the independent approach in terms of deviance. However, there may be a plausible explanation for the observed results: Like many other cognitive models (e.g., Brown, Neath, & Chater, 2007; Nosofsky, 1986; van Ravenzwaaij et al., 2011; Wetzels et al., 2010), TAX and CPT were implemented using a probabilistic choice rule (Eq. 5). Within this choice rule, a sensitivity parameter (θ, in our case) governs how deterministically the option with the higher valuation is chosen (here, higher parameter values lead to increasingly deterministic choices). Therefore, as long as the decision maker chooses the option with the higher valuation more than 50 % of the time (which is mostly the case for the Glöckner & Pachur, 2012, data), higher values of θ will always yield a higher proportion of correct predictions. As we mentioned above, due to parameter shrinkage, the hierarchical model led to smaller θ values on the group level than did the arithmetic mean of the independent estimates; this, in turn, produced less deterministic predictions, and hence a lower proportion of correct predictions.

To illustrate the relationship between deviance and posterior predictive accuracy, we ran a simulation (using TAX) in which we estimated both the posterior predictive accuracy and the deviance for predicting the observed choices at t2 for different values of the θ parameter (keeping the other parameters constant at the group-level estimates obtained at t1). Figure 8 displays the results. The figure shows a nonmonotonic relationship between θ and deviance: Deviance is lowest (indicating the highest likelihood) right around the group-level posterior mean for θ at 0.14, and sharply increases for both smaller and larger values of θ. The reason why the minimum is located near the mean of the posterior of t2 is that in the context on hand, the deviance, which exerts a high influence on the posterior distribution, had a similar shape at t1 and t2. In contrast to the shape of the deviance curve, Figure 8 further shows that the proportion of correct predictions monotonically increases with higher values of θ. This simulation illustrates that although posterior predictive probability and deviance are related (since both are calculated from the probability that the model correctly predicts the data), the two measures can systematically diverge. If a model assigns a high likelihood to most observed data (yielding a high posterior predictive probability) but strongly errs for a single observed datum (i.e., assigns an extremely low likelihood to it), the product of the likelihoods will be low, and hence deviance will escalate (see Selten, 1998).

General discussion

Bayesian hierarchical approaches to parameter estimation are based on the principle of partial pooling, in which estimates on the individual level are informed by group-level distributions. One of the advantages of this approach is that it can lead to more reliable parameter estimates than can independent techniques. We compared hierarchically and independently estimated parameters in order to evaluate two prominent models of decision making under risk, CPT and TAX, with respect to different aspects of model generalizability. The results of this comparison, drawing on an empirical data set obtained by Glöckner and Pachur (2012), indicated that the relative performance of the hierarchical approach varied across the different aspects of model generalizability. Specifically, the hierarchical approach led to higher parameter stability across time, as measured by the coefficient of variation, and to higher predictive accuracy, as measured by deviance. By contrast, the hierarchical approach did not consistently increase—and sometimes even decreased—the test–retest correlations of the parameters, and it led

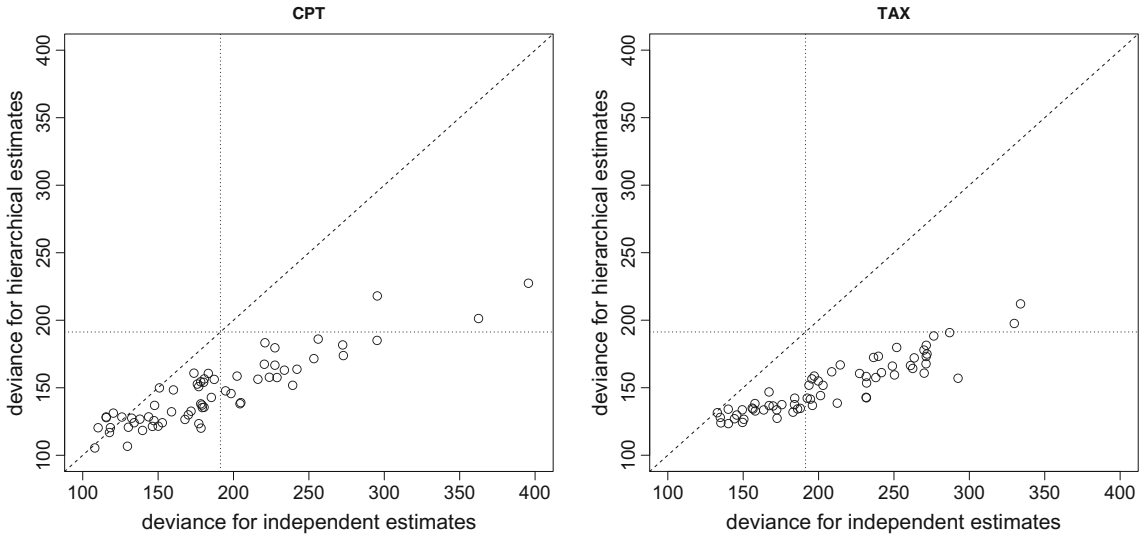

Fig. 7 Mean deviances for individual choices at t1, averaged across representative samples of the posterior distributions for CPT (left plot) and TAX (right plot), estimated at t1. Each dot indicates one participant. Points below the diagonal indicate lower deviance (hence, better fit) for the hierarchically estimated parameters over the independently estimated parameters. The dotted lines indicate the deviance for random predictions

to a lower posterior predictive accuracy. These results held for both CPT and TAX.

One reason for the finding that the hierarchical approach did not produce superior results on all measures of generalizability was that partial pooling induced a high amount of shrinkage for the group-level distributions, such that extreme yet reliable parameter estimates were strongly "pulled in" toward the group-level mean. In the context of the models investigated here, this shrinkage also led to a lower estimate of the choice sensitivity parameter, and hence to less deterministic predictions. As we have shown, this in turn led to lower posterior predictive accuracy.

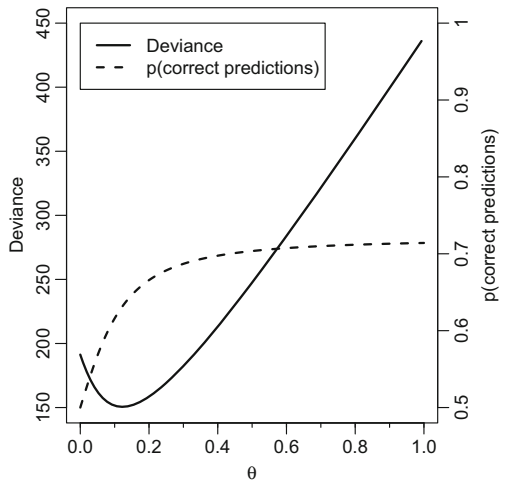

Fig. 8 Relationship between the θ parameter and deviance (solid line, left y-axis) and the mean proportion of correct predictions across all participants (dashed line, right y-axis) for the TAX model

Importantly, however, the strong shrinkage in the hierarchical approach presumably was fueled by strong interdependencies between some of the model parameters—in particular, the sensitivity parameter in the choice rule and the curvature parameter of the models' utility functions. Thus, the high shrinkage does not seem to be a drawback or flaw of the hierarchical approach itself, but rather points to identification issues with the model parameters. Next, we discuss the implications of this result in more detail. This is followed by a discussion of further factors that may contribute to the differences observed between the hierarchical and independent approaches.

Strong parameter interdependence occurs in many cognitive models

One may suspect that the strong parameter interdependence that we observed, and that seems to be a key factor for the differences between the independent and hierarchical approaches, may be a peculiarity of the models investigated here, CPT and TAX. However, several other cognitive models of decision making feature a similar parameterization with a probabilistic choice rule and a utility function. For some of them, past research has also found considerable correlations between these parameters, such as for models for the Bayesian analog risk task (van Ravenzwaaij et al. 2011) and the expectancy valence model (and variants thereof; Wetzels et al., 2010). In fact, simulation results by Stewart (2011) suggest that strong interdependence between parameters should be expected in any model that has a probabilistic choice rule and a utility function with adjustable parameters. Parameterized choice

rules combined with parameterized item-strength functions also exist in other areas of cognitive science, such as categorization and memory (e.g., Brown et al., 2007; Nosofsky, 1986). It will therefore be interesting to see to what extent the results that we obtained would also occur for hierarchical implementations of these models.

In general, such correlations between model parameters are disadvantageous: They make it difficult to precisely estimate the parameters and interpret them in isolation, and they suggest that the models might benefit from a reparameterization to eliminate redundancies. Ways to reparameterize the models would be to express one parameter as a function of another or to set some parameters to fixed values. For instance, instead of a parameterized utility function, a simple identity function [i.e., $u(x) = x$] could be used. Alternatively, in some cases it might be possible to employ a nonparameterized choice rule, such as the original version of Luce's choice rule (e.g., Stott, 2006). However, if researchers are genuinely interested in the value of that parameter, these solutions are not helpful. If parameter interdependence cannot be avoided or is theoretically justified, and researchers aim to test and compare hierarchical models, a pragmatic solution could be to implement correlated prior distributions in the first place (e.g., Pratte & Rouder, 2011).

As we will illustrate next, even if parameters are interdependent, the advantages of the hierarchical approach will often prevail—in particular, when there are very few data on the individual level and when very little is known about the possible range of the parameter values (as reflected in wide prior distributions).

Modeling with sparse data on the individual level

A particular advantage of hierarchical over independent approaches is the ability to yield reliable estimates even when relatively few data are available for individuals (e.g., Busemeyer & Diederich, 2010). With sparse data, estimating parameters using the independent approach is rather error-prone, and the principle of borrowing strength implemented in hierarchical techniques can increase accuracy. In the present case, a relatively large number of observations (138) were available per individual, allowing quite reliable estimates using the independent approach. To explore whether in the case on hand the hierarchical approach plays out its strengths even in terms of posterior predictive accuracy when only sparse data are available on the individual level, we conducted an additional analysis in which we determined the predictive accuracy of TAX on the basis of only 10 % (selected randomly) of the original 138 choices for each participant at t1. To retain the proportions of the different types of lottery problems, we drew seven gain, three loss, and four mixed lotteries. For this subset of data, we then estimated the TAX parameters using both the independent and hierarchical approaches. From the estimates, predictive accuracy was assessed for predictions across time (i.e., predicting each individual's choices at t2 on the basis of the respective individual parameter estimates at t1) and for out-of-sample predictions (i.e., predicting individual choices at t2 on the basis of the group-level mean at t1).

For individual predictions across time, the hierarchical approach yielded a posterior predictive accuracy that was about 4 % higher than that yielded by the independence approach. This difference was statistically different from zero, as indicated by a conventional t test, $t(63) = 4.1$, $p = .0001$, $BF_{10} = 152$, indicating that the hierarchical approach now outperformed the independent approach. For out-of-sample predictions, the posterior predictive accuracy for the hierarchical estimates ($p = .66$, $SD = .03$) was similar to that of the independent estimates ($p = .66$, $SD = .06$). Thus, in this latter case, the hierarchical approach was no longer inferior to the independent approach (but also no better). Overall, these results suggest that one factor contributing to the relatively strong performance of the independent approach in our analyses was that the estimation error was already quite low, due to the large amount of data available for each individual.

Using extremely uninformative priors

The posterior predictive accuracy for the out-of-sample predictions reported in the previous section may appear quite high, given the very small number of observations that informed these estimates. One reason for this may be that some informed assumptions about the underlying behavioral patterns were already built into the structures of the models themselves; for instance, the assumed range of the priors was not completely uninformative, but was based on theoretical reasoning and previous research. In fact, even when simply drawing random samples from the prior distributions (i.e., not taking any observations into account), TAX already achieved a posterior predictive accuracy of .56, which is significantly better than chance, $t(63) = 11.6$, $p < .001$, $BF_{10} > 10{,}000$. However, when the ranges of the uniform priors were increased by a factor of 20 (i.e., ranging from 0 to 100 for θ, γ, and β, and from –50 to 50 for δ), the predictive posterior accuracy dropped to .53 [although this was still better than chance: $t(63) = 5.85$, $p < .001$, $BF_{10} > 10{,}000$]. This indicates that the assumed range of priors had an impact on the models' predictive accuracies.

To examine whether using these less informative, wide prior ranges also affected the relative predictive accuracies of hierarchical versus independent parameter estimates, we again estimated the TAX parameters from a random subset of 10 % of the observed choices, but now using the wider range of priors. In this case, the posterior predictive accuracy when predicting new group members eventually exceeded that of

the independent approach (p = .63, SD = .03, vs. p = .58, SD = .07), $t(63) = 4.8$, $p < .001$, $BF_{10} = 1{,}571$.[9]

Is sum, these results suggest that hierarchical estimation techniques are particularly advantageous when only very few observations are available for each individual and when there is little prior knowledge—even for a model that shows considerable parameter interdependence. In these situations, independent estimates will yield very vague and uninformative posterior distributions, and borrowing strength from the estimations of other individuals is particularly helpful in order to position the posterior distributions within reasonable bounds.

The role of group-level priors

Given the integral role of priors in Bayesian modeling, one may suspect that an additional factor contributing to the high amount of shrinkage observed in our main analyses was the choice of priors on the group level. What speaks against this explanation for the present case is that the priors spanned a rather wide range of possible values, including those reported in previous studies (e.g., Glöckner & Pachur, 2012; Rieskamp, 2008). Furthermore, there was no indication for any of the parameters that the estimated posterior distributions (including the group-level variances) were squeezed to the respective prior boundaries, and the priors thus seemed to be sufficiently wide. Arguably, priors should reflect reasonable expectations in light of what is currently known about a model (Edwards, Lindman, & Savage, 1963), and this was followed in our implementations of the models. As we mentioned above, our analyses indicated that the shrinkage diminished considerably when we removed the correlation between parameters in the model. Nevertheless, an alternative choice of priors or different model specifications might have induced a more appropriate degree of shrinkage, and hence a greater advantage of hierarchical over independent estimation techniques, even when parameters were correlated.

Hierarchical modeling and the reduction of error variance

In a classic article, Efron and Morris (1977) showed theoretically that hierarchical approaches that induce shrinkage of individual estimates toward the grand mean often lead to smaller prediction errors than do independent estimates. Is our result that hierarchical estimation sometimes performs worse than independent estimation at odds with this finding? In Efron and Morris's analyses, predictive accuracy was measured as the sum across the squared estimation errors, a measure that is closely related to the deviance criterion that we used.[10] When we quantified predictive accuracy for the Glöckner and Pachur (2012) data in terms of squared estimation error, our results resembled those of Efron and Morris, since we also found an advantage of the hierarchical over the independent modeling approach: On average (across all participants), the estimation error for the hierarchical parameter estimates was smaller (i.e., better) by a factor of 1.16 (TAX) and 1.1 (CPT), as compared with the independent approach.

Capturing the data-generating process

In general, if a given model can capture the processes that give rise to the observed data, all unexplained variance has to be due to measurement error. In this case, hierarchically informed estimates are particularly well suited to reduce prediction errors (Efron & Morris, 1977). However, when empirical choice data are analyzed, the data-generating process is usually unknown, and the observed variance in the data is not just due to error, but also stems from systematic influences that the model does not capture. In this case, the advantage of consolidating information through hierarchical modeling will be reduced (Efron & Morris, 1977). In the present context—where the goal was to model the outcome of a presumably complex cognitive process, and the actual data-generating process was unknown—it is difficult to assess the degree to which the models captured systematic or unsystematic influences. To be sure, both models considered in our analyses—CPT and TAX—are highly successful accounts of decision making and have been shown to account for a large number of phenomena and to perform well across a wide range of situations (e.g., Birnbaum, 2008; Glöckner & Pachur, 2012). However, given the complexity of human decision making, it seems plausible that these models could nevertheless provide only an approximation of the true data-generating processes (e.g., Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013; Su et al., 2013). Therefore, deviations between the models' predictions and observed data are presumably not just due to random error, but may also reflect systematic divergences between the models and the actual cognitive processes. This issue of possible model misspecification might pose a challenge to hierarchical approaches, since it could potentially lead to shrinkage of both random error and systematic bias, even when the latter is robust (and should therefore not be shrunk).

Further advantages of hierarchical parameter estimation

Although our results point to some situations in which hierarchical techniques may not consistently improve model

[9] The range of the prior distribution has very little impact on the results when taking all 138 choices into account. Presumably, with this amount of data on the individual level, the influence of the prior on the posterior estimates is negligible.

[10] If p is a vector of probabilities for making a correct prediction, the deviance is defined as $-2*\text{sum}[\log(p)]$, whereas the squared error is defined as $\text{sum}[(1-p)^2]$.

generalizability in terms of posterior predictive accuracy or test–retest correlations, we would like to highlight that there are also other reasons to use a hierarchical rather than an independent modeling approach. As was indicated by our results, when the criterion was deviance, squared estimation error, or the coefficient of variation, the hierarchical estimates were clearly superior. We also found that, even when parameters are correlated, shrinkage will increasingly help to reduce unsystematic noise when there are few data for each individual and very little prior information. In these situations, independent estimates are often not feasible, so hierarchical approaches are advisable, even if the goal is to increase posterior predictive accuracy (Busemeyer & Diederich, 2010).

Furthermore, hierarchical approaches naturally lend themselves to comparing groups as a whole and to quantifying the variability between individuals (Lee & Webb, 2005). As our examples show, they also provide a principled way to capture correlations between parameters on the group level, and thus offer crucial insights that would otherwise be difficult to obtain (Lee & Wagenmakers, 2014). Last but not least, as was pointed out by Lee and Newell (2011), "one of the most compelling features of the hierarchical Bayesian approach is that it encourages deeper theorizing and the construction of more psychologically complete models" (pp. 839–840), which will help to advance our understanding of the cognitive processes underlying behavior.

References

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26,* 217–238.

Berkowitsch, N. A. J., Scheibehenne, B., & Rieskamp, J. (2014). Testing multialternative decision field theory rigorously against random utility models. *Journal of Experimental Psychology: General, 143,* 1331–1348. doi:10.1037/a0035159

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115,* 463–501. doi:10.1037/0033-295X.115.2.463

Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes, 71,* 161–194.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327,* 307–310.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539–576. doi:10.1037/0033-295X.114.3.539

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. New York, NY: Sage.

Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging, 28,* 64–76. doi:10.1037/a0029875

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

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

Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision, 60,* 283–313.

Fox, C. R., & Poldrack, R. A. (2008). Prospect theory and the brain. In P. W. Glimcher, E. Fehr, C. Camerer, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 145–174). San Diego, CA: Academic Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition, 123,* 21–32. doi:10.1016/j.cognition.2011.12.002

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94,* 236–254. doi:10.1037/0033-295X.94.2.236

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38,* 129–166.

Harbaugh, W. T., Krause, K., & Vesterlund, L. (2002). Risk attitudes of children and adults: Choices over small and large probability gains and losses. *Experimental Economics, 5,* 53–84.

Hendricks, W. A., & Robey, K. W. (1936). The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics, 7,* 129–132.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30,* 1–15.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. doi:10.1177/1745691611406925

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making, 6,* 832–842.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12,* 605–621. doi:10.3758/BF03196751

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 720–738. doi:10.1037/a0022639

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Li, S.-C., Lewandowsky, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General, 125,* 360–369. doi:10.1037/0096-3445.125.4.360

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28,* 3049–3067.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology, 55,* 84–93. doi:10.1016/j.jmp.2010.08.006

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, 28, 924–940. doi:10.1037/0278-7393.28.5.924

Pachur, T., Hanoch, Y., & Gummerum, M. (2010). Prospects behind bars: Analyzing decisions under risk in a prison population. *Psychonomic Bulletin & Review, 17,* 630–636. doi:10.3758/PBR.17.5.630

Pachur, T., Hertwig, R., Gigerenzer, G., & Brandstätter, E. (2013). Testing process predictions of models of risky choice: A quantitative model comparison approach. *Frontiers in Psychology, 4,* 646. doi:10.3389/fpsyg.2013.00646

Pachur, T., Hertwig, R., & Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision, 1,* 64–78.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology, 65,* 207–240. doi:10.1016/j.cogpsych.2012.03.003

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 1–10. Retrieved from www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology, 55,* 36–46. doi:10.1016/j.jmp.2010.08.007

Qiu, J., & Steiger, E.-M. (2011). Understanding the two components of risk attitudes: An experimental analysis. *Management Science, 57,* 193–199.

R Development Core Team. (2012). *R: A language and environment for statistical computing [Computer software]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1446–1465. doi:10.1037/a0013646

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12,* 573–604. doi:10.3758/BF03196750

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General, 137,* 370–389. doi:10.1037/0096-3445.137.2.370

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review, 120,* 39–64. doi:10.1037/a0030777

Scheibehenne, B., & Studer, B. (2014). A hierarchical Bayesian model of the influence of run length on sequential predictions. *Psychonomic Bulletin & Review, 20,* 211–217. doi:10.3758/s13423-013-0469-1

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136,* 414–429. doi:10.1037/0096-3445.136.3.414

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics, 1,* 43–62.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284.

Stewart, N. (2011). Information integration in risky choice: Identification and stability. *Frontiers in Psychology, 2,* 301. doi:10.3389/fpsyg.2011.00301

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32,* 101–130.

Su, Y., Rao, L.-L., Sun, H.-Y., Du, X.-L., Li, X., & Li, S. (2013). Is making a risky choice based on a weighting and adding process? An eye-tracking investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1765–1780. doi:10.1037/a0032861

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

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

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology, 55,* 94–105. doi:10.1016/j.jmp.2010.08.010

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology, 54,* 14–27. doi:10.1016/j.jmp.2008.12.001

Yechiam, E., & Busemeyer, J. R. (2008). Evaluating generalizability and parameter consistency in learning models. *Games and Economic Behavior, 63,* 370–394.

Yechiam, E., & Ert, E. (2011). Risk attitude in decision making: In search of trait-like constructs. *Topics in Cognitive Science, 3,* 166–186.