Mem Cogn (2016) 44:454–468
DOI 10.3758/s13421-015-0567-6

Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats

Martha Michalkiewicz[1] · **Edgar Erdfelder**[1]

Published online: 16 November 2015

Abstract The recognition heuristic (RH) is a simple decision strategy that performs surprisingly well in many domains. According to the RH, people decide on the basis of recognition alone and ignore further knowledge when faced with a recognized and an unrecognized choice object. Previous research has revealed noteworthy individual differences in RH use, suggesting that people have preferences for using versus avoiding this strategy that might be causally linked to cognitive or personality traits. However, trying to explain differences in RH use in terms of traits presupposes temporal and cross-situational stability in use of the RH, an important prerequisite that has not been scrutinized so far. In a series of four experiments, we therefore assessed the stability in RH use across (1) time, (2) choice objects, (3) domains, and (4) presentation formats of the choice objects. In Experiment 1, participants worked on the same inference task and choice objects twice, separated by a delay of either one day or one week. Experiment 2 replicated Experiment 1 using two different object sets from the same domain, whereas Experiment 3 assessed the stability of RH use across two different domains. Finally, in Experiment 4 we investigated stability across verbal and pictorial presentation formats of the choice objects. For all measures of RH use proposed so far, we found strong evidence for both temporal and cross-situational stability in use of the RH. Thus, RH use at least partly reflects a person-specific style of decision making whose determinants await further research.

Keywords Decision making · Individual differences · Cognitive trait · Multinomial processing tree models · Hierarchical Bayesian modeling

Which city is more populous: Tokyo or Busan? If you recognize Tokyo but not Busan, you can use a simple inference strategy: the fast-and-frugal recognition heuristic (RH; Goldstein & Gigerenzer, 2002). According to the RH, a person should choose the recognized object and ignore any further knowledge. Thus, when following the RH, you would choose Tokyo simply because you recognize it. Alternatively, you can deliberately integrate knowledge available over and above recognition—for instance, that Tokyo has an international airport and that cities with an international airport are (most often) more populous. In this case, you would arrive at the same conclusion with both decision strategies. However, which factors are responsible for using the RH versus integrating further knowledge?

There is a large body of research on the situational determinants of RH use. In general, RH use increases, the greater the importance of a quick decision (Hilbig, Erdfelder, & Pohl, 2012; Pachur & Hertwig, 2006) and the higher the validity of the recognition cue (Castela, Kellen, Erdfelder, & Hilbig, 2014; Hilbig, Erdfelder, & Pohl, 2010; Pachur, Mata, & Schooler, 2009; Pohl, 2006; Scheibehenne & Bröder, 2007). By contrast, integration of further knowledge increases as knowledge becomes more easily available and easier to integrate (Bröder & Eichler, 2006; Glöckner & Bröder, 2011; Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015;

Electronic supplementary material The online version of this article (doi:10.3758/s13421-015-0567-6) contains supplementary material, which is available to authorized users.

✉ Martha Michalkiewicz
michalkiewicz@psychologie.uni-mannheim.de

[1] Department of Psychology, School of Social Sciences, University of Mannheim, Schloss, Ehrenhof-Ost, 68131 Mannheim, Germany

Hilbig, Pohl, & Bröder, 2009; Newell & Fernandez, 2006; Richter & Späth, 2006). In sum, it is quite well established that participants adjust their RH use according to situational factors (for reviews, see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011; Pohl, 2011).

However, studies that have addressed situational factors have also revealed large individual differences in RH use (Hilbig & Richter, 2011; Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2010; Newell & Shanks, 2004; Pachur, Bröder, & Marewski, 2008). As was argued by Gigerenzer and Brighton (2009, p. 133), "in virtually every task we find individual differences in strategies."

Figure 1 displays the individual proportions of RH use on the basis of data from Hilbig and Pohl (2009, Exp. 1), assessed with the *r*-model (Hilbig et al., 2010) that we will describe in detail below. Why would RH use differ to such a degree between participants under constant context conditions? As is indicated by the standard errors illustrated in Fig. 1, the observed heterogeneity is too large to be attributable to error variance only. In fact, a goodness-of-fit test of the homogeneity hypothesis that all 24 individual proportions of RH use are equal reveals a clear misfit [$\Delta G^2(23) = 72.8$, $p < .001$]. This suggests that the heterogeneity might reflect individual preferences for certain strategies that are not determined by the context. This, in turn, gives rise to the question: Do individual traits underlie RH use?

Indeed, there is some evidence that different groups of people prefer different strategies. In particular, Pachur et al. (2009) showed that elderly people use the RH more often than young adults do (see also Horn, Pachur, & Mata, 2015). Extending this line of research to the life span, Pohl, von Massow, and Beckmann (2015) detected a nonmonotonic trend in RH use in younger age groups: Preadolescent children and young adults used the RH about equally often, whereas adolescents used it more frequently. Moreover, exploring a different source of individual differences, Hilbig and Pohl (2008) found that more knowledgeable people tend to rely less on the RH. These examples show that groups of individuals may differ significantly from each other in RH use even if the decision context is kept constant. In addition, at least one study successfully examined the relationship between RH use and personality traits: Hilbig (2008) demonstrated that neuroticism was positively related to RH use. Results like this one encourage a search for traits as sources of individual differences in RH use.

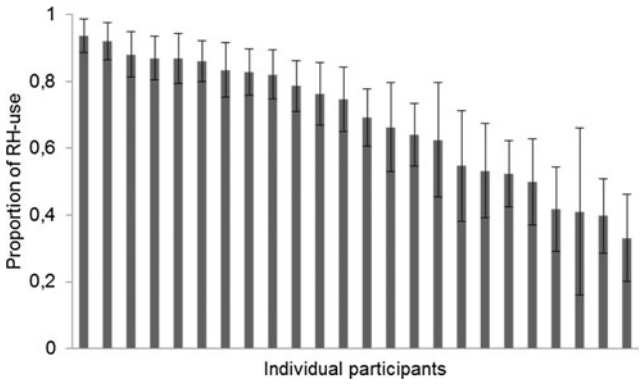

Fig. 1 Individual proportions of RH use per participant for the data from Hilbig and Pohl (2009, Exp. 1). RH use is estimated via the *r* parameter of the *r*-model (Hilbig et al., 2010) using multiTree (Moshagen, 2010) and ordered by size. Error bars illustrate standard errors.

However, considerable evidence also shows how difficult it is to find associations between strategy use and individual traits. For instance, apart from neuroticism, Hilbig (2008) also investigated agreeableness, conscientiousness, openness, and extraversion, but did not find any substantial effect on RH use. Similarly, Pachur et al. (2009) tested associations between measures of inhibitory control and RH use without finding evidence for substantial correlations. Furthermore, Bröder (2012) summarized multiple studies on the take-the-best heuristic (TTB; Gigerenzer & Goldstein, 1999) and various cognitive and personality traits. Only intelligence was found to affect adaptive use of the TTB heuristic, depending on the environmental payoff structure (see also Bröder, 2003). Notably, none of the remaining variables covered by Bröder's (2012) review, including need for cognition, impulsivity, and the Big Five personality traits, showed any substantial relation.

Given the equivocal evidence on traits as determinants of decision strategy use, we argue that prior to trying to explain variability in RH use in terms of cognitive or personality traits, an important precondition should be checked: stability in use of the RH. The need for stability assessment has previously been pointed out by Bröder and Newell (2008, p. 208): "It is yet an open question whether the different strategy preferences diagnosed in a one-shot assessment of an experiment will turn out to be stable across tasks and situations." If RH use turns out to be stable, it makes sense to search for relations between RH use and individual traits. If, in contrast, stability cannot be shown—that is, if RH use varies haphazardly within individuals across situations, then it will hardly be possible to find replicable relations between RH use and cognitive or personality traits. Of course, stability does not imply that each individual behaves identically in all situations—that is, exhibits exactly the same level of RH use everywhere. Rather, it means that individual behavior is "meaningfully consistent" (Roberts, 2009, p. 139)—for example, that participants showing higher than average RH use in one decision context will also tend to show higher than average RH use in a second context.

Notably, several studies have already emphasized stability in behaviors related to judgment and decision making. Thus, evidence suggesting the stability of RH use would fit nicely into related lines of research. For instance, Witkin,

Goodenough, and Karp (1967) showed that the ability to ignore the visual context in perceptual judgments—known as *field independence*—is a stable cognitive style from childhood to young adulthood, at least. Furthermore, Couch and Keniston (1960) investigated acquiescence bias in terms of consistency over time and generality over tests. More recently, Kantner and Lindsay (2012, 2014) found response bias in recognition tasks to be stable across time, stimulus materials, and item presentation formats. Similarly, Aminoff and colleagues (2012) demonstrated stability in criterion shifting in recognition memory across different presentation formats. Furthermore, Odum (2011) reanalyzed prior studies on delay discounting, and the results suggested stability across time for up to one year using many different stimulus materials.

Recently, the focus has been shifted to systematic tests of parameter stability of cognitive models. For instance, Yechiam and Busemeyer (2008) evaluated the stability of several learning models for repeated choice problems across different tasks. Also, Glöckner and Pachur (2012) examined parameter stability with respect to cumulative prospect theory across time (see also Scheibehenne & Pachur, 2015). These are just some examples of stability research, and many more could be listed.

Following a route similar to those in the studies outlined above, we will analyze four different aspects of stability—namely, stability across (1) time, (2) choice objects, (3) domains, and (4) presentation formats. For this purpose, we conducted four experiments in which participants completed two sets of inference tasks. Stability was measured as the test–retest correlation between RH use in Tests 1 and 2. In Experiment 1, we assessed stability across time for delays of one day versus one week, using exactly the same choice objects in both tests. Experiment 2 was designed to replicate Experiment 1. This time, however, different choice objects were drawn from the same domain on the two tests, providing for an assessment of stability across disjoint sets of objects. To further analyze the influence of the choice materials, we conducted Experiment 3, in which we assessed the stability of RH use across different domains. Finally, to examine stability across presentation formats, we designed Experiment 4, in which we used names versus pictures as formats of object presentation.

Note that, in all four experiments, we specifically opted to investigate stability factors that are unconfounded with the overall level of RH use. Investigating the same people twice—even when using different materials or presentation formats—is usually assumed not to affect the overall level of RH use, provided that the recognition and knowledge validities do not differ between tests. To the degree that we succeeded in implementing these conditions, we expected neither main effects of situational factors nor interaction effects with individual factors on RH use, enabling us to assess the influence of individual differences without confounds.

General method

When investigating stability in use of the RH, it is important to ensure that both strategies under consideration—the RH and knowledge use—are applicable in the current context. For this purpose, several conditions must be fulfilled. First, participants must recognize at least one, and at most all but one, of the objects in order to apply the RH in the first place. Optimally, participants should recognize half of the objects to maximize the proportion of cases in which the RH can be applied. Second, participants must obviously have some kind of knowledge about the set of objects and the question of interest. If nothing but recognition information is available, participants obviously cannot apply more elaborated strategies incorporating further knowledge. Third, the validity of the recognition cue α (i.e., the proportion of cases in which choosing the recognized object leads to a correct response) and the validity of knowledge β (i.e., the proportion of cases in which the application of knowledge leads to a correct response) should both be greater than chance. These conditions render both the RH and the use of knowledge reasonable strategies. When selecting the materials for our experiments, we aimed at satisfying all of these requirements.

To analyze our data, we primarily relied on the *r*-model (Fig. 2; Hilbig et al., 2010), a multinomial processing tree model (Batchelder & Riefer, 1999; Erdfelder et al., 2009) tailored to measure RH use, as defined by Goldstein and Gigerenzer (2002). Specifically, if exactly one object is recognized, participants will either apply the RH with probability r or make use of further knowledge with probability $1 - r$. We focused on this model because it successfully decontaminates the probability of RH use from the effects of knowledge-based strategies that might also lead to choice of the recognized object (see Hilbig, 2010).

In the present analyses, we applied the latent-trait approach to multinomial processing tree models (Klauer, 2010) because it elegantly handles variability in parameters between individuals. For this purpose, we constructed a hierarchical version of the *r*-model (Fig. 3) based on the implementation by Matzke, Dolan, Batchelder, and Wagenmakers (2015) and extended it to account for the data of two test occasions simultaneously. Compared to standard correlational analyses, the latent-trait approach has one main advantage (Klauer, 2010; Matzke et al., 2015): It allows for the joint estimation of model parameters and the correlations between parameters in a single step. The estimated correlations are thus automatically adjusted for the uncertainty in the individual parameter estimates; that is, the model estimates the correlation of the true scores decontaminated from error influences. For a comprehensive introduction to hierarchical models and their advantages, see, for instance, Lee and Wagenmakers (2013).

Extension of the single-test hierarchical *r*-model to a two-test version is straightforward. We estimated the parameters of this extended model within the Bayesian framework using Markov chain Monte Carlo sampling employing OpenBUGS

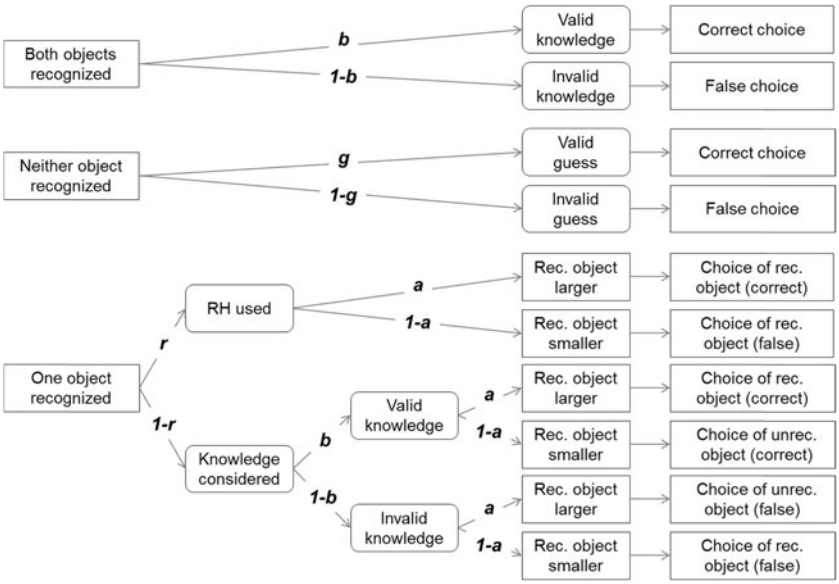

Fig. 2 Illustration of the *r*-model. Displayed are the three object pairs that can occur in a paired-comparison task: knowledge pairs (both objects recognized), guessing pairs (neither object recognized), and recognition pairs (exactly one object recognized). Participants' responses are assigned to eight categories and accounted for by four latent parameters—namely, recognition validity (parameter *a*), knowledge validity (parameter *b*), the probability of correct guessing (parameter *g*), and most importantly, the probability of RH use (parameter *r*). Adapted from "One-Reason Decision Making Unveiled: A Measurement Model of the Recognition Heuristic," by Hilbig et al. 2010, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, p. 125. Copyright 2010 by the American Psychological Association.

(Lunn, Spiegelhalter, Thomas, & Best, 2009) through R2WinBUGS (Sturtz, Ligges, & Gelman, 2005).[1] For each analysis, we ran three chains with 500,000 iterations each, using a thinning rate of 10, and discarded the first 100,000 iterations as a burn-in period. Chain convergence was reached for all estimated parameters ($R < 1.01$; Gelman, Carlin, Stern, & Rubin, 2004). Also, the effective sample sizes were sufficient to trust the parameter estimates (Kruschke, 2014).[2] As is common practice in Bayesian analysis, we will report the means of the posterior distributions together with their 95 % Bayesian credible intervals (BCI). In particular, μ^{r1} and μ^{r2} are interpreted as the group-level estimates of RH use on Test Occasions 1 and 2, respectively. On the basis of the covariance matrix Σ, the standard deviations σ^{r1} and σ^{r2} as well as the correlations between parameters $\rho_{r1,r2}$ are derived. In this context, standard deviations reflect the variation between participants, being close to zero when participants are rather homogeneous and large when there are substantial individual differences.

Experiment 1

To assess stability over time, we tested whether a given participant would show similar levels of RH use for a set of choice objects on two different points in time.

[1] The R code, the model file, and a sample data set are provided in the online supplemental materials.

[2] To meet these criteria, we increased the number of iterations per chain to 1 million for the week group of Experiments 1 and 2 as well as for the different group of Experiment 3.

Method

Design and procedure To study the RH, we employed the frequently used city-size task (Goldstein & Gigerenzer, 2002; Hilbig & Pohl, 2009; Pachur et al., 2008) including, in random order, a paired-comparison task and a recognition task. In the comparison task, participants were asked to decide for pairs of cities which of the two cities was more populous. In the recognition task, participants had to indicate for all cities whether or not they had heard of the city before the experiment.

To test our hypothesis, participants were randomly assigned to one of two groups and worked on the city-size task twice, separated by a delay of either one day (*day group*) or one week (*week group*). To render the procedures equivalent for both groups, all participants completed three sessions—the initial session and the two following sessions one day and one week later—at exactly the same time of day. In Session 1, all participants first worked on the city-size task and then completed two unrelated experiments to render the intention of the study less obvious. In Sessions 2 and 3, depending on the group, participants either worked on the city-size task for the second time or again completed an unrelated experiment. To control for contamination of the results, we asked participants two questions after they had completed the city-size task for the second time: (1) whether they had tried to memorize their responses in the first session, and (2) whether they had looked up city sizes after the first session. Participants received course credit or a flat fee of €10 after (and only after) completion of all three sessions. We refrained from using performance-contingent payment here, as

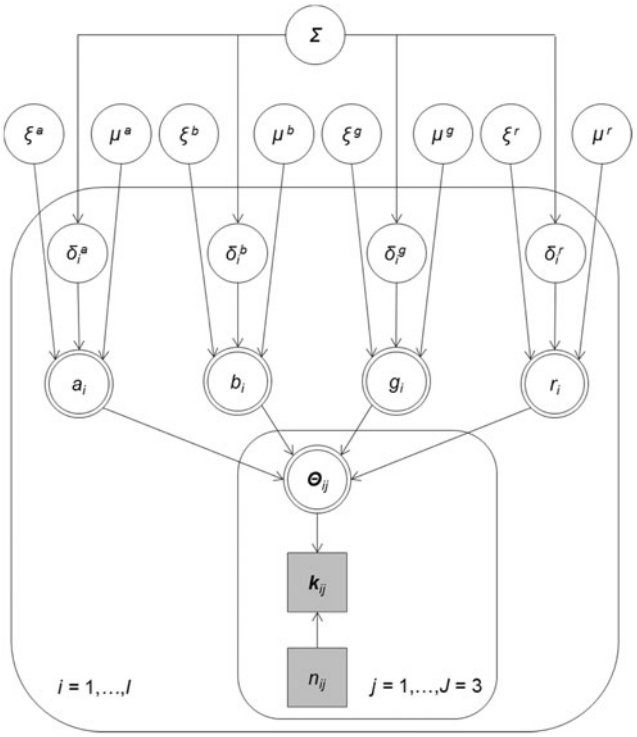

Fig. 3 Illustration of the single-test hierarchical latent-trait *r*-model (cf. Matzke, Lee, & Wagenmakers, 2013). Displayed are the relations between data and latent model parameters. Observable and unobservable variables are presented as shaded versus unshaded nodes, respectively; discrete and continuous variables are presented as squares versus circular nodes. To-be-estimated and derived variables (which can be parameterized by means of the remaining model parameters) are presented as single-bordered versus double-bordered nodes. The plates indicate replications over I individuals for the $J = 3$ object cases (knowledge, guessing, and recognition cases). For each individual i and each object case j, the vector of category counts $\mathbf{k}_{ij}$ follows a multinomial distribution with probability vector $\mathbf{\Theta}_{ij}$ and number of observations n_{ij}, as is presented in the *r*-model (see Fig. 2). The individual model parameters s_i ($s \in \{a, b, g, r\}$) are modeled in a probit-transformed space, $s_i \leftarrow \Phi(\mu^s + \xi^s * \delta_i^s)$, as linear combinations of the group-level mean $\mu^s \sim N(0, 1)$, the multiplicative scale parameter $\xi^s \sim U(0, 100)$, and the individual displacement parameter δ_i^s, drawn from a common multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, with $\mathbf{\Sigma}^{-1}$ following a Wishart($\mathbf{I}$, 5) distribution.

is often done in this kind of task, because we were concerned that this would encourage participants to look up the city sizes between the first and second tests of the RH.

Material We used exactly the same objects in both repetitions of the city-size task. Specifically, we selected a random sample of 100 cities from the 150 most populous US cities for the recognition tasks. From these items, we randomly created one sample of 300 pairs for the comparison tasks, ensuring that (1) each city appeared exactly six times and (2) recognition and knowledge validity were adequate to render both RH use and knowledge use reasonable strategies to solve the task. To achieve this, we selected the materials on the basis of the data of pilot experiments on RH use conducted in our lab[3].

Participants A total of 70 student participants were recruited via posters and mailing lists at the University of Mannheim. Six participants dropped out of the experiment by not coming back to the second or third session. The remaining 64 participants consisted of 45 women and 19 men, between 18 and 30 years of age ($M = 21.8$ years, $SD = 2.7$). All participants were native speakers or fluent in German.

Results and discussion

The recognition proportions, recognition validities, and knowledge validities did not differ significantly between groups [all ts(62) < 1.67, ps > .10, Bayes factors (BF_{10}s) < 0.82].[4] Therefore, the analyses related to these variables are reported for both groups combined. As expected, the mean recognition validity[5] $\overline{\alpha}$ and the mean knowledge validity[6] $\overline{\beta}$ showed that both strategies—the RH and knowledge use—were clearly better than guessing ($\overline{\alpha} = .77$, $SD = .06$; $\overline{\beta} = .67$, $SD = .08$, and $\overline{\alpha} = .77$, $SD = .06$; $\overline{\beta} = .66$, $SD = .08$, for Tests 1 and 2, respectively) [all ts(63) > 15.0, ps < .001, BF_{10}s > 1,000]. Hence, the application of either of these strategies was reasonable. Furthermore, participants in the initial session recognized about half of the cities ($M = 49.1$ objects, $SD = 11.4$), resulting in a sufficient number of recognition cases. Surprisingly, participants recognized more cities on the second test of the RH ($M = 58.6$ objects, $SD = 19.8$) than on the first [$t(63) = 4.71$, $p < .001$, $BF_{10} > 1{,}000$]. This was most probably due to confusion of real recognition (i.e., cities seen before the experiment) and familiarity induced by the presentation of the cities in the initial session. Hence, since the recognition judgments collected on the second test were obviously biased, we based all measures on the recognition judgments obtained in the first test only. However, using the original recognition judgments of each session did not change the results substantially (see Table 2 in the Appendix).

To further control for possible confounds, we analyzed the control questions. All participants confirmed that they had not looked up the city sizes. This was validated by the numbers of correct answers in the comparison tasks across the two tests:

[3] The stimulus materials of all experiments are provided in the online supplemental materials.

[4] Bayes factors were computed using the BayesFactor R package (Rouder, Speckman, Sun, Morey, & Iverson, 2009) and interpreted following the classification by Jeffreys (1961).

[5] Recognition validity is computed as the proportion of recognition pairs where the recognized object represents the correct choice. It matches the estimates for parameter a of the hierarchical r-model.

[6] Knowledge validity is computed as the proportion of correct choices in knowledge pairs. It matches the estimates for parameter b of the hierarchical *r*-model.

Indeed, we even found a slight decrease (M = 201.9 answers, SD = 12.5, and M = 199.0 answers, SD = 13.1, for Tests 1 and 2, respectively) [$t(63) = 2.66$, $p = .01$, $BF_{10} = 3.48$]. Moreover, only nine participants stated that they had tried to memorize their answers during the initial session. Obviously, memorizing the answers to 300 questions, given only 1,351 ms per question (i.e., the average response time across participants and items in the comparison task), and retaining them for up to one week is very unlikely. The actual answers confirmed this: Participants decided differently on the second test than on the first for up to half of the trials (M = 69.0, SD = 18.8, Min = 38, Max = 142). In sum, there was no indication that judgments in the second comparison task were biased.

As expected, we found strong heterogeneity in RH use between participants (μ^{r1} = .77 [.74, .80], σ^{r1} = .38 [.29, .48], and μ^{r2} = .70 [.66, .74], σ^{r2} = .42 [.33, .53], for Tests 1 and 2, respectively). Unexpectedly, we observed a small decrease in the r parameter between the two test occasions (Δ_r = .07 [.04, .10]), showing that participants used the RH slightly less often in the second test than in the first. However, this difference should not have influenced the core results, because the correlation between RH use on Tests 1 and 2 is independent of the mean level of RH use.

To test the main hypothesis, we examined the correlations between the r parameter on Tests 1 and 2 for both groups separately. Overall, we observed strong positive correlations for both groups ($\rho_{r1,r2}$ = .80 [.56, .94], and $\rho_{r1,r2}$ = .71 [.39, .91], for the day and week groups, respectively). The small drop in correlations after an extended delay of one week was not reliable, as can be seen by the overlapping BCIs. This provided evidence in favor of our hypothesis that people used the RH consistently across time.

To establish a benchmark against which to compare the sizes of the correlations, we estimated the *within-test correlation* of the r parameter for single tests. More precisely, we split the data of each participant into the first and the last 150 trials. We then estimated the correlations between the parameters of these two parts using the hierarchical r-model. The magnitude of consistency across tests was similar to that observed within a single test ($\rho_{r1,r2}$ = .73 [.46, .91] on Test 1 and $\rho_{r1,r2}$ = .72 [.47, .89] on Test 2), showing that the delay between task repetitions (0 h, 24 h, or 168 h) had little effect on the test–retest correlation. In sum, the results reflect stability across time up to one week, at least when using the same choice objects repeatedly.

A possible objection against Experiment 1 is that participants perhaps just behaved very similarly when working on exactly the same task and choice objects twice. Why should they change their judgments when facing the same choice objects for the second time, perhaps even remembering (some of) the choices they had made previously? To test whether stability was caused by the invariance of materials only, we conducted a second experiment using different objects in the two inference tasks.

Experiment 2

Method

Design and procedure The design and procedure were identical to those of Experiment 1. Participants were again randomly assigned to one of two groups and worked on the city-size task twice: in the initial session and either one day (*day group*) or one week (*week group*) after the initial session.

Material This time, two disjoint samples of 25 cities were randomly drawn from the 61 most populous world cities for the two recognition tasks. Each of the 25 cities was exhaustively paired, resulting in two samples of 300 pairs for the two comparison tasks. Because we wanted the two iterations of the city-size task to closely resemble each other except for object identity, we selected the materials on the basis of the data of prior experiments (Hilbig et al., 2010, 2012). Thereby, we made sure that both city samples had similar proportions of recognized objects, recognition, and knowledge validities.

Participants A total of 94 student participants were recruited via posters and mailing lists at the University of Mannheim. Six participants completed the first session only, and thus dropped out of the experiment. This resulted in 88 participants, consisting of 49 women and 39 men, between 18 and 59 years of age (M = 22.7 years, SD = 5.3). All participants were native speakers or fluent in German.

Results and discussion

Five participants had to be excluded from the analyses because they recognized all of the objects or all but one. Because group differences were again negligible [all ts(81) < 1.21, ps > .23, BF_{10}s < 0.43], the analyses of recognition rates, recognition validities, and knowledge validities were based on the data of both groups combined. First, participants recognized on average 17.3 cities (SD = 2.8) of Set 1 and 16.5 cities (SD = 1.8) of Set 2, resulting in sufficient recognition cases for both sets. Second, the mean recognition and knowledge validities showed that both strategies—the RH and further knowledge—were appropriate—that is, better than guessing ($\overline{\alpha}$ = .69, SD = .08; $\overline{\beta}$ = .57, SD = .08, and $\overline{\alpha}$ = .68, SD = .07; $\overline{\beta}$ = .61, SD = .08, for Sets 1 and 2, respectively) [all ts(82) > 7.61, ps < .001, BF_{10}s > 1,000]. In sum, the materials were selected in line with our goals.

Replicating Experiment 1, the r parameter, representing the proportion of RH use, showed strong variability between participants (μ^{r1} = .59 [.51, .66], σ^{r1} = .86 [.71, 1.03], and μ^{r2} = .73 [.66, .78], σ^{r2} = .82 [.69, .99], for Tests 1 and 2, respectively). Again, we observed a difference in the average r

parameters between the two choice sets, this time opposite to that in Experiment 1: Participants used the RH on average *less* often in the first than in the second test (Δ_r = –.14 [–.20, –.07]). As in Experiment 1, this difference should have impacted stability only marginally.

To test the main hypothesis, we again examined the correlations between the first and second tests of the RH for both groups separately. The results showed very similar positive correlations ($\rho_{r1,r2}$ = .54 [.26, .75] and $\rho_{r1,r2}$ = .53 [.25, .75] for the day and week groups, respectively). This suggests that RH use is stable across time even when the choice objects differ in the two tests, ruling out the objection to Experiment 1 that stability is perhaps limited to exact replications of choices.

Notably, the between-test correlations in Experiment 2 were lower than the within-test correlations ($\rho_{r1,r2}$ = .90 [.80, .96] and $\rho_{r1,r2}$ = .91 [.83, .97] for the first and second tests, respectively). The former also tended to be lower than the between-test correlations observed in Experiment 1. This suggests that consistency in RH use partly depends on the similarity of (or overlap in) choice objects. To further study the influence of differences in materials, we conducted a third experiment in which we assessed the stability of RH use across *different* domains.

Experiment 3

Method

Design and procedure Participants worked on two tests of the RH (both consisting of a recognition task and a comparison task, in random order) within a single session. As in Experiment 2, we used different materials on the two tests of the RH. However, there was one important modification: The corresponding materials were drawn from two different judgment domains. The experiment again comprised two groups. In the *related group*, participants worked on two different object sets drawn from similar (although not identical) domains. In the *different group*, in contrast, the two object sets were drawn from clearly distinct domains. To maintain a high level of motivation across the lengthy experiment, participants received performance-contingent payment. In both comparison tasks, participants gained €0.03 for each correct judgment, whereas they lost €0.03 for each false judgment. However, to avoid strategy-learning effects, participants received feedback about their performance at the end of the experiment only.

Material We used different materials for the two groups. In the *related* group, participants were asked to decide on (1) the success of celebrities and (2) the success of movies, in random order. The domain of celebrities consisted of the 100 most successful celebrities according to the Forbes List 2012 (www.forbes.com), which defined success as entertainment-related earnings plus media visibility. The domain of movies contained the 100 most successful German movies, characterized by the numbers of cinemagoers in Germany. Analogously, in the *different* group, participants were asked to decide on (1) the size of islands and (2) the success of musicians, in random order. The domain of islands included the 60 largest islands worldwide, whereas the domain of musicians involved the world's 150 most successful musicians, characterized by the numbers of records sold worldwide. To make sure that for both groups the two choice sets had similar properties (i.e., proportions of objects recognized, recognition, and knowledge validities), objects were selected on the basis of the recognition judgments of an independent prestudy. Specifically, we chose a random sample of 25 objects for each of the four domains for the recognition task. Each of these samples was then exhaustively paired, resulting in 300 trials for the comparison task.

Participants A total of 135 student participants were recruited at the University of Mannheim and randomly assigned to one of the two groups outlined above. The sample consisted of 87 women and 48 men, between 18 and 45 years of age (M = 21.6 years, SD = 3.6). All participants were native speakers or fluent in German. They received an average salary of €3.70 (SD = 1.85).

Results and discussion

Three participants had to be excluded from the analyses because they recognized either all but one or none of the objects. Descriptive analyses revealed that the materials were chosen in line with our goals: Participants recognized on average 13.8 celebrities (SD = 3.3) and 15.4 movies (SD = 4.0), as well as 12.6 islands (SD = 2.6) and 16.0 musicians (SD = 4.0), resulting in sufficient numbers of recognition cases for all domains. In the *related* group, the mean recognition and knowledge validities did not differ across materials ($\overline{\alpha}$ = .64, SD = .10; $\overline{\beta}$ = .56, SD = .09, and $\overline{\alpha}$ = .64, SD = .09; $\overline{\beta}$ = .56, SD = .10, for celebrities and movies, respectively) [all ts(67) < 0.35, ps > .72, BF_{10}s < 0.14). In the *different* group, the mean recognition validities were similar for both domains ($\overline{\alpha}$ = .68, SD = .08, and $\overline{\alpha}$ = .69, SD = .14, for islands and musicians, respectively) [t(63) = 0.64, p = .53, BF_{10} = 0.17]. The difference in mean knowledge validities was most probably due to the choice of the materials being based on a small prestudy ($\overline{\beta}$ = .65, SD = .08, and $\overline{\beta}$ = .60, SD = .08, for islands and musicians, respectively) [t(63) = 3.94, p < .001, BF_{10} = 109.0]. Overall, the recognition and knowledge validities showed that both strategies—RH use and knowledge use—were reasonable [all ts(67) > 4.63, ps < .001, BF_{10}s > 1,000 for the *related* group; all ts(63) > 10.5, ps < .001, BF_{10}s > 1,000 for the *different* group].

Again, we found strong heterogeneity in RH use between participants in both groups (μ^{r1} = .77 [.72, .82], σ^{r1} = .68 [.56, .83], and μ^{r2} = .83 [.78, .86], σ^{r2} = .63 [.50, .78], for celebrities and movies, respectively; μ^{r1} = .69 [.61, .76], σ^{r1} = .84 [.68, 1.04], and μ^{r2} = .82 [.76, .87], σ^{r2} = .76 [.62, .95], for islands and musicians, respectively). As before, the difference in mean levels of RH use should not have influenced the results in a crucial manner (Δ_r = –.05 [–.10, .001] and Δ_r = –.13 [–.21, –.05], for the *related* and *different* groups, respectively).

To test the main hypothesis, the correlations between RH use in the two tasks were examined separately for each group. Overall, the results showed medium to strong correlations for both groups ($\rho_{r1,r2}$ = .42 [.18, .62] and $\rho_{r1,r2}$ = .33 [.08, .55], for the *related* and *different* groups, respectively), thus supporting the hypothesis of stability across domains. However, both correlations were substantially lower than the within-test correlations ($\rho_{r1,r2}$ = .81 [.66, .92] for celebrities and $\rho_{r1,r2}$ = .77 [.59, .90] for movies; $\rho_{r1,r2}$ = .89 [.79, .96] for islands and $\rho_{r1,r2}$ = .81 [.65, .92] for musicians). These differences demonstrate a potential impact of variation in domains. Also, there seems to be a downward trend in the stability coefficients, as compared to the results of Experiments 1 and 2.

In conclusion, the results of Experiments 1 to 3 support the hypothesis that RH use is relatively stable when using the same or different choice objects and also when using similar or clearly distinct domains. However, they also support the conjecture that the overall similarity of the to-be-compared decision scenarios impacts the stability of decision strategies across domains.

Another important aspect of the choice context has not been addressed so far: Experiments 1 to 3 presented choice options in verbal form only. This is a rather abstract presentation format relative to typical choice situations in everyday life. Does stability in RH use generalize to perceptually enriched, and presumably ecologically more valid, pictorial presentations of choice objects? Experiment 4 was designed to address this question.

Experiment 4

Method

Design and procedure Participants worked on two choice tasks within one session using exactly the same materials but different presentation formats. In the first task, the choice options were indicated verbally (i.e., using names), whereas in the second task they were indicated pictorially (i.e., using photos), or vice versa. Once again, participants received performance-contingent payment to maintain a high motivational level throughout the experiment, but they were informed about their overall performance only after the whole experiment was completed.

Material We used the names and pictures of the 100 most successful celebrities according to the Forbes List 2012 as choice options (cf. Exp. 3). For the recognition task, a subset of 25 objects was randomly chosen without repetition. This set was exhaustively paired, resulting in 300 pairs of objects for the comparison task. To guarantee similar properties of the materials, we selected the objects on the basis of the recognition judgments of an independent prestudy.

Participants A total of 87 student participants were recruited via posters and mailing lists at the University of Mannheim. The sample consisted of 58 women and 29 men, between 18 and 45 years of age (M = 22.3 years, SD = 4.7). All participants were native speakers or fluent in German, and they received an average salary of €3.19 (SD = 1.36).

Results and discussion

Descriptive analyses revealed that the materials had been chosen in line with our goals: Participants recognized on average about half of the objects (M = 13.2 celebrities, SD = 3.7, presented as names, and M = 12.5 celebrities, SD = 3.7, presented as pictures), resulting in sufficient recognition cases. Similarly, the actual mean recognition and knowledge validities showed that both the RH and knowledge use were reasonable strategies under both presentation formats ($\overline{\alpha}$ = .64, SD = .10; $\overline{\beta}$ = .62, SD = .09, and $\overline{\alpha}$ = .65, SD = .10; $\overline{\beta}$ = .58, SD = .11, for names and pictures, respectively) [$t(86) > 6.75$, $p < .001$, $BF_{10} > 1{,}000$].

As before, we found large individual differences in RH use, irrespective of the presentation format (μ^{r1} = .74 [.68, .79], σ^{r1} = .83 [.70, 1.00], and μ^{r2} = .71 [.64, .77], σ^{r2} = .84 [.70, 1.00], for names and pictures, respectively; Δ_r = .03 [–.03, .09]). More importantly, RH use was stable across presentation modes, demonstrated by a strong positive correlation of $\rho_{r1,r2}$ = .60 [.44, .74]. The within-test correlations were very similar and considerably higher for each format ($\rho_{r1,r2}$ = .91 [.83, .96] and $\rho_{r1,r2}$ = .89 [.80, .95], for names and pictures, respectively), relative to the correlations across presentation formats. Thus, stability in RH use appears to depend, at least to a certain extent, on invariance of the presentation formats.

Stability of alternative measures of RH use

To make sure that evidence on within-individual stability was not tied to a particular measure (cf. Kantner & Lindsay, 2012) or to statistical peculiarities of the *r*-model that might bias stability assessment, we replicated the main analyses using all measures of RH use previously employed in the relevant literature: the adherence rate (i.e., the proportion of cases in which the recognized object is chosen), the indices *c* (i.e., the tendency to follow

the recognition cue) and d' (i.e., the ability to discriminate cases in which recognition yields a correct vs. a false inference) derived from signal detection theory (Pachur et al., 2009), and the discrimination index (DI; Hilbig & Pohl, 2008), similar to the discriminability parameter d'. For this purpose, we calculated the respective measures for each participant and each test occasion separately and used standard methods of stability assessment (i.e., Pearson product-moment correlation coefficients), summarized in Table 1.

Briefly, we found the same pattern of results as with the *r*-model. In fact, the correlation coefficients for the adherence rate and the index *c* were comparable in size to those for the *r* parameter.[7] However, the correlation coefficients for the indices DI and d' are somewhat smaller. These indices have in common that they capture the ability to discriminate cases in which the RH leads to correct versus false inferences. As such, they measure the deviation from pure RH use. Uncontrolled noise factors—for instance, the overall degree of knowledge about the domain—might affect the degree of deviation from perfect RH use. Therefore, neither the lack of a linear relationship between the *r* parameter and these two indices (Horn et al., 2015) nor the lower within-test correlations (see Table 2 in the Appendix) and, by implication, the lower stability of DI and d' come as a surprise.

General discussion

When making decisions, people can use different strategies: These include simple strategies like the fast-and-frugal recognition heuristic (Goldstein & Gigerenzer, 2002), which assumes that decisions are based on recognition exclusively, and more costly strategies such as the integration of knowledge stored in memory, which demand more time and cognitive resources. There are two general approaches to identifying factors that influence strategy selection. On the one hand, a fertile line of research focuses on external factors—that is, situational and domain-specific variables (e.g., Bröder & Eichler, 2006; Hilbig et al., 2010; Newell & Shanks, 2004; Richter & Späth, 2006). On the other hand, a sparsely studied line of research has focused on internal factors, such as personality traits and other persistent individual characteristics (e.g., Hilbig, 2008; Pachur et al., 2009).

It has been shown repeatedly that people differ to a large extent in applying specific strategies. This heterogeneity appears to be caused by person-specific factors, independent of contextual influences (e.g., Gigerenzer & Brighton, 2009; Hilbig & Richter, 2011; Pachur et al., 2008). Among others, Shiloh, Koren, and Zakay (2001, p. 701) observed that "individuals seem to have personal tendencies that favor the use of compensatory or non-compensatory decision strategies, which are based on personality traits." Consequently, research on personality influences is important because contextual aspects alone cannot explain individual differences in strategy selection satisfactorily. However, prior to exploring the personality determinants of RH use, temporal and cross-situational stability in RH use needs to be demonstrated as an important precondition. If people do not apply the RH in a consistent way, it will eventually turn out to be impossible to find replicable relations between RH use and individual traits.

For these reasons, we conducted four experiments to assess four different aspects of stability in RH use—namely, stability across (1) time, (2) choice objects, (3) domains, and (4) presentation formats. In all four experiments, participants worked on two tasks measuring RH use. The stability of RH use was assessed as the cross-task correlation. To account for measurement and sampling errors in individual parameter estimates of RH use, we used a hierarchical extension of the *r*-model applied to the two test occasions (Klauer, 2010; Matzke et al., 2015). Moreover, to ensure that the results were not limited to a particular measure of RH use, we also evaluated stability for all measures of RH use previously employed in the relevant literature (see Hilbig, 2010, for a review and comparative evaluation).

In Experiments 1 and 2, we assessed stability across time using a delay of either one day or one week between the two choice tasks. The only difference between the two experiments was that we used exactly the same choice objects in both tasks in Experiment 1 and different choice objects in Experiment 2. The results of both experiments confirmed our hypothesis that RH use is stable across time. To provide a benchmark against which to compare the correlation coefficients, we estimated the *within-test correlation*. To this end, we split the data into the first and second 150 trials and estimated the correlations in RH use between these two parts for both test occasions separately. This coefficient can be interpreted as the "baseline" stability[8] for a zero delay. In Experiment 1, the correlations of both groups were comparable to the within-test correlations in Tests 1 and 2. In Experiment 2, the correlations across test occasions were somewhat lower than the within-test correlations. However, the within-test correlations of Experiment 2 were based on exactly the same stimulus materials, whereas the between-test correlations were based on two distinct material sets. In sum, these findings suggest that stability is largely unaffected by the time

[7] The smaller correlations in the day group of Experiment 2 compared to the week group are due to a single participant. Excluding this participant resulted in correlation coefficients of .60 for the adherence rate and of .57 for the index *c*.

[8] The within-test correlations are similar in size across Experiments 2–4, where each object was repeated 24 times in the decision task. A slight decrease was found for Experiment 1, where each object was repeated only six times. This small number of repetitions might have caused a difference in choice objects between the first and the second halves of the decision task, resulting in a somewhat smaller within-test correlation than in the experiments in which the same objects were repeated 24 times.

Table 1 Pearson correlation coefficients (and 95 % confidence intervals) across the two tests of the recognition heuristic (RH) for all measures of RH use previously employed in the relevant literature

Exp.	Group	Measure of RH Use			
		Adherence Rate	*c*	*d′*	DI
1	Day group	.90 [.81, .95]	.86 [.73, .93]	.47 [.15, .70]	.39 [.05, .65]
	Week group	.72 [.49, .86]	.70 [.46, .84]	.51 [.19, .73]	.52 [.20, .74]
2	Day group	.45 [.17, .67]	.48 [.20, .69]	.12 [–.19, .41]	.42 [.13, .64]
	Week group	.55 [.30, .73]	.55 [.30, .73]	.004 [–.30, .31]	.23 [–.08, .50]
3	Related group	.49 [.28, .65]	.48 [.27, .65]	.05 [–.19, .29]	.05 [–.19, .29]
	Different group	.43 [.21, .61]	.46 [.24, .63]	.28 [.04, .49]	.20 [–.05, .42]
4	All participants	.64 [.50, .75]	.63 [.48, .74]	.32 [.12, .50]	.27 [.06, .45]

DI, discrimination index. The correlation coefficients are shown for Experiments 1–4, separately for experimental groups.

interval between measurement occasions, but might be influenced by differences in stimulus materials.

To study stability across the different materials, we systematically increased the differences in choice objects and domains in Experiments 1 to 3. Participants worked repeatedly on exactly the same choice objects in Experiment 1, on different choice objects drawn from the same domain in Experiment 2, and on objects from two distinct judgment domains that differed slightly versus substantially in Experiment 3. Medium to strong stability in RH use was found in all experiments. However, the stability coefficients tended to decrease when the differences between tasks increased: The correlations were very high and comparable to the within-test correlations when using exactly the same objects in both tests. They were lower when using different stimulus materials from the same object domain, dropped again when using objects from slightly different domains, and dropped even more when using objects from substantially different domains.

Finally, we examined stability across different presentation formats in Experiment 4, in which we presented the choice objects as names versus pictures. As before, RH use was relatively stable across tasks. However, as compared to the within-test correlations (which were similar for the two presentation formats and in line with those in the other experiments), stability was reduced slightly by a change in presentation formats.

One potential objection refers to the possibility that the observed stability in RH use was perhaps nothing but an epiphenomenon of stability in the participants' knowledge. Hilbig and Pohl (2009) have shown that more valid knowledge leads to less use of the RH. Therefore, one might hypothesize that the stability of RH use was caused by stable underlying differences in knowledge validity, with individuals high in knowledge validity using the RH less often than those with low knowledge validity. However, recall that we found stability across different domains in Experiment 3. Assuming that knowledge validity is domain-specific (i.e., uncorrelated between domains), cross-domain stability in RH use indicates that this result cannot be accounted for solely by stability in knowledge validity. Of course, one could maintain that some aspects of knowledge are perhaps domain-general, leading to positive knowledge correlations between domains. For instance, people who have more valid knowledge concerning the domain of celebrities might also have more valid knowledge concerning movies. However, at least three aspects of our results are inconsistent with the idea that stability of RH use is caused by individual differences in knowledge (see Tables 4 to 6 in the Appendix). First, we found a reliable negative correlation between RH use and knowledge validity for one group in Experiment 2 only. Second, the retest correlations between knowledge validities either were comparable in size to the retest correlations for RH use or were even smaller and not reliable. Third, the correlations between RH use on Tests 1 and 2 were very similar in size to the partial correlations (partialing out the effect of knowledge measured on Test 1 or 2, respectively), showing that the stability of RH use is unaffected by individual differences in knowledge.

Arguing along similar lines, one might hypothesize that stability in RH use is perhaps an epiphenomenon of stability in individual recognition validities. It has repeatedly been shown that recognition validity differences between domains are positively correlated with domain-specific RH use (Gigerenzer & Goldstein, 2011; Hilbig et al., 2010; Pachur et al., 2011). However, we found no reliable positive relation between individual RH use and the individual recognition validities in any experiment (see Table 4). Also, stability in RH use is unaffected by individual differences in recognition validities, since the partial correlations closely match the zero-order test–retest correlations (see Table 6). Note that this result is not in direct conflict with the previous studies cited above, because these studies assessed recognition validity differences *between domains*, whereas we investigated recognition validity differences *between individuals*.

In sum, stability in RH use was found across time, choice objects, domains, and presentation formats to a degree similar to what has previously been found for some other trait-like

variables in judgment and decision making.[9] Moreover, the stability of RH use is not affected by individual differences in knowledge or recognition validities, suggesting that it truly reflects a specific style of decision making rather than individual differences in the information on which decisions and inferences are based.

Stability as an important precondition opens the way for exploring personality as a source of individual variation in decision-making styles. However, our results also reveal one limitation: We should not expect correlations between RH use and personality traits larger than the stability coefficients observed here. If the correlation between two tasks measuring RH use in different domains does not exceed .33 (Exp. 3, *different* group), then the correlation between a powerful personality predictor and RH use should not be expected to exceed this value, either. The insight that even powerful predictors can be expected to show moderate correlations at best provides a possible explanation for the difficulties in finding replicable relations between RH use and individual traits (Hilbig, 2008; Pachur et al., 2009).

Furthermore, our work also opens the way for exploring another rather neglected influence on decision-making styles: the *interaction* of personality and situational factors. For instance, Bröder (2003) showed that intelligence moderates adaptive use of the TTB heuristic, depending on whether or not TTB performs well in a given decision context. In our view, an analogous effect of intelligence on adaptive RH use is worth investigating. One could also think of other potential interaction effects. Hilbig (2008), for instance, suggested that participants high in neuroticism prefer RH use over knowledge use in order to avoid a diagnostic test of their abilities. If this holds, increasing the self-value relevance of the task might boost the effect of neuroticism. Furthermore, certain personality traits possibly reveal their influence only under certain situational conditions. For instance, even if impulsivity by itself is not a predictor of strategy selection (Bröder, 2012), a context condition such as time pressure might turn it into one. By contrast, strong situational influences might also eliminate the effect of personality. For instance, the lack of evidence for an association between strategy use and the need for cognition (Bröder, 2012) might originate from situational influences overshadowing personality influences. Controlling for situational influences as strictly as possible might reveal that the need for cognition is indeed an important predictor. We thus suggest using strictly neutral decision contexts (i.e., "weak situations"; cf. Mischel, 1973) if the goal is to study pure influences of personality traits on strategy use.

Moreover, following Kantner and Lindsay's (2012, 2014) analysis of individual differences in response bias, we might ask whether RH use can be conceived of as a *cognitive trait*, meaning "an aspect of cognition that typifies an individual" (Kantner & Lindsay, 2012, p. 1164). Given the present data, we cannot answer this question now. However, we are sure that it will inspire future research. It would be interesting to analyze, for example, whether people who prefer RH use over knowledge use also favor other fast-and-frugal heuristics, such as the TTB heuristic (Gigerenzer & Goldstein, 1999). Given that the recognition and TTB heuristics share the one-reason decision-making principle, correlations between preferences for the two heuristics seem very likely. Furthermore, one could also explore whether RH use is related to other response tendencies and biases as part of "a more general, intra-individually stable decision-making heuristic" (Kantner & Lindsay, 2012, p. 1175).

In any case, one important conclusion can be drawn from the present study: The likelihood of RH use is not only influenced by situational determinants that affect the costs and benefits of RH use (Erdfelder, Küpper-Tetzel, & Mattern, 2011; Hilbig et al., 2010; McCloy, Beaman, Frosch, & Goddard, 2010; Oppenheimer, 2003; Pachur & Biele, 2007; Pohl, Erdfelder, Hilbig, Liebke, & Stahlberg, 2013; Schooler & Hertwig, 2005). As we have shown in the present research, it is also influenced by relatively stable individual tendencies favoring either RH use or the integration of further knowledge. Thus, our work contributes to a new line of research on the cognitive and personality traits underlying RH use, aiming at a comprehensive theory that integrates situational and personality determinants of decision strategies.

Author note We thank Benjamin Hilbig and Rüdiger Pohl for providing the raw data of Hilbig and Pohl (2009). We are also thankful to Michael Lee, Thorsten Pachur, and an anonymous reviewer for their thoughtful comments on a previous version of the manuscript. The research reported in this article was supported by grants from the German Research Foundation (DFG; Grant Nos. ER 224/2-1 and ER 224/2-2). Parts of this research were presented at the Conference of Experimental Psychologists (Wien, 2013; Gießen, 2014).

Appendix

Table 2 Main results concerning RH use in Experiment 1, using the original recognition judgments of each session

	Day Group		Week Group	
	Test 1	Test 2	Test 1	Test 2
Mean	.76 [.71, .80]	.73 [.66, .79]	.78 [.73, .83]	.69 [.62, .75]
SD	.39 [.27, .55]	.48 [.34, .67]	.39 [.27, .56]	.43 [.30, .60]
Correlation	.67 [.34, .88]		.58 [.23, .84]	

Means, standard deviations, and correlation coefficients (with 95 % Bayesian credible intervals) are measured via the hierarchical *r*-model and shown separately for experimental groups and the two tests of the RH. *Test 1* refers to the initial test of the RH, whereas *Test 2* refers to the second test of the RH, done one day or one week later.

[9] To illustrate, a stable response bias was shown using different time intervals ($\rho \in [.67, .73]$) and presentation formats ($\rho \in [.33, .81]$; Kantner & Lindsay, 2012). Similarly, delay discounting was found to be stable across time ($\rho \in [.71, .91]$), stimulus materials ($\rho \in [.18, .90]$), and presentation formats ($\rho \in [.44, .83]$; Odum, 2011).

Table 3 Within-test correlation coefficients (and 95 % confidence intervals) for all measures of RH use previously employed in the relevant literature

Exp.	Material	Measure of RH Use			
		Adherence Rate	*c*	*d′*	DI
1	Test 1	.80 [.69, .87]	.79 [.67, .86]	.07 [–.18, .31]	.33 [.09, .53]
	Test 2	.84 [.75, .90]	.76 [.63, .84]	.40 [.18, .59]	.40 [.18, .59]
2	Test 1	.91 [.86, .94]	.91 [.86, .94]	.50 [.32, .64]	.63 [.47, .74]
	Test 2	.90 [.85, .93]	.91 [.87, .94]	.65 [.50, .76]	.77 [.66, .84]
3	Celebrities	.88 [.81, .92]	.88 [.82, .93]	.20 [–.04, .42]	.21 [–.03, .43]
	Movies	.82 [.72, .88]	.80 [.69, .87]	.38 [.15, .57]	.31 [.08, .51]
	Islands	.93 [.89, .96]	.92 [.87, .95]	.38 [.15, .57]	.49 [.28, .66]
	Musicians	.86 [.78, .91]	.87 [.79, .92]	.47 [.25, .64]	.58 [.39, .73]
4	Names	.92 [.89, .95]	.93 [.89, .95]	.46 [.27, .61]	.44 [.26, .60]
	Pictures	.91 [.87, .94]	.91 [.86, .94]	.55 [.38, .68]	.55 [.38, .68]

DI, discrimination index. The correlation coefficients are shown for Experiments 1–4, separately for the two tests of the RH. For Experiments 1 and 2, *Test 1* refers to the initial test of the RH, whereas *Test 2* refers to the second test of the RH, performed one day or one week later. Within-test correlations are estimated using the Spearman–Brown-corrected Pearson correlation coefficient.

Table 4 Correlation coefficients (with 95 % Bayesian credible intervals) for the correlations between RH use and recognition validity and between RH use and knowledge validity, separately for the two tests of the RH

Exp.	Group	Correlation Between RH Use and			
		Recognition Validity		Knowledge Validity	
		Test 1	Test 2	Test 1	Test 2
1	Day group	.35 [–.09, .71]	.37 [–.05, .71]	.13 [–.05, .55]	.29 [–.14, .68]
	Week group	.41 [–.01, .73]	.31 [–.12, .67]	.25 [–.48, .80]	.38 [–.13, .79]
2	Day group	–.06 [–.40, .29]	–.16 [–.48, .19]	**–.35 [–.62, –.02]**	**–.36 [–.63, –.04]**
	Week group	–.01 [–.10, .35]	.06 [–.32, .42]	–.03 [–.16, .31]	.01 [–.33, .35]
3	Related group	–.02 [–.28, .25]	.16 [–.11, .42]	.36 [.09, .60]	.07 [–.22, .36]
	Different group	.23 [–.05, .48]	.22 [–.03, .46]	.19 [–.11, .48]	.19 [–.12, .47]
4	All participants	.01 [–.22, .25]	–.19 [–.40, .40]	.58 [.36, .76]	.22 [–.01, .45]

The correlation coefficients are shown for Experiments 1–4, separately for experimental groups and the two tests of the RH. Boldface indicates significant correlations, shown by credible intervals that do not include 0. For Experiments 1 and 2, *Test 1* refers to the initial test of the RH, whereas *Test 2* refers to the second test of the RH, performed one day or one week later. For Experiment 3, *Tests 1 and 2* refer to the two different domains that were used as materials (celebrities vs. movies and islands vs. musicians). For Experiment 4, *Tests 1 and 2* refer to the two presentation formats (names vs. pictures). Recognition and knowledge validities are assessed via the *a* and *b* parameters of the *r*-model, respectively. Correlations are estimated using the hierarchical *r*-model.

Table 5 Correlation coefficients (with 95 % Bayesian credible intervals) for recognition and knowledge validities across tests

Exp.	Group	Test-Retest Correlations	
		Recognition Validity	Knowledge Validity
1	Day group	.81 [.57, .95]	.67 [.29, .91]
	Week group	.88 [.71, .97]	.48 [–.24, .88]
2	Day group	–.10 [–.47, .27]	.53 [.21, .78]
	Week group	.05 [–.08, .43]	.51 [.41, .78]
3	Related group	.38 [.11, .62]	.06 [–.28, .38]
	Different group	.43 [.18, .65]	.28 [–.10, .61]
4	All participants	.16 [–.08, .38]	.70 [.46, .85]

The correlation coefficients are shown for Experiments 1–4, separately for experimental groups. Recognition and knowledge validities are assessed via the *a* and *b* parameters of the *r*-model, respectively. Correlations are estimated using the hierarchical *r*-model.

Table 6 Comparison between zero-order and partial correlation coefficients (with 95 % Bayesian credible intervals) for RH use across tests, partialing out the effects of the recognition validities of Tests 1 and 2 and the knowledge validities of Tests 1 and 2, respectively

Exp.	Group	Zero-Order Correlation	Partial Correlation Controlling for			
			Recognition Validity		Knowledge Validity	
			Test 1	Test 2	Test 1	Test 2
1	Day group	.80 [.55, .94]	.78 [.52, .94]	.78 [.54, .94]	.81 [.57, .95]	.79 [.55, .94]
	Week group	.71 [.39, .91]	.66 [.33, .89]	.66 [.32, .89]	.66 [.24, .90]	.66 [.30, .90]
2	Day group	.54 [.26, .75]	.55 [.28, .75]	.54 [.26, .75]	.48 [.19, .71]	.46 [.16, .69]
	Week group	.53 [.25, .75]	.53 [.24, .76]	.55 [.28, .77]	.53 [.25, .75]	.54 [.25, .76]
3	Related group	.42 [.17, .63]	.42 [.18, .62]	.40 [.15, .61]	.36 [.10, .59]	.42 [.17, .63]
	Different group	.33 [.08, .55]	.33 [.08, .55]	.34 [.09, .56]	.28 [.003, .52]	.29 [.03, .52]
4	All participants	.60 [.43, .74]	.60 [.44, .74]	.59 [.43, .73]	.48 [.26, .67]	.57 [.39, .72]

The zero-order and partial correlation coefficients are shown for Experiments 1–4, separately for experimental groups. For Experiments 1 and 2, *Test 1* refers to the initial test of the RH, whereas *Test 2* refers to the second test of the RH, performed one day or one week later. For Experiment 3, *Tests 1 and 2* refer to the two different domains that were used as materials (celebrities vs. movies and islands vs. musicians). For Experiment 4, *Tests 1 and 2* refer to the two presentation formats (names vs. pictures). Recognition and knowledge validities are assessed via the *a* and *b* parameters of the *r*-model, respectively. Correlations and partial correlations are estimated using the hierarchical *r*-model.

References

Aminoff, E. A., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., . . . Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*, 1016–1030. doi:10.3758/s13421-012-0204-6

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86. doi:10.3758/BF03210812

Bröder, A. (2003). Decision making with the "Adaptive Toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 611–625. doi:10.1037/0278-7393.29.4.611

Bröder, A. (2012). The quest for take the best—Insights and outlooks from experimental research. In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 216–240). New York, NY: Oxford University Press.

Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica, 121,* 275–284. doi:10.1016/j.actpsy.2005.07.001

Bröder, A., & Newell, B. R. (2008). Challenging some common beliefs: Empirical work within the adaptive toolbox metaphor. *Judgment and Decision Making, 3,* 205–214. **Retrieved from http://journal.sjdm.org/bn2.pdf**

Castela, M., Kellen, D., Erdfelder, E., & Hilbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review, 21,* 1131–1138. doi:10.3758/s13423-014-0587-4

Couch, A., & Keniston, K. (1960). Yeassayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60,* 151–174. doi:10.1037/h0040372

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Journal of Psychology, 217,* 108–124. doi:10.1027/0044-3409.217.3.108

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making, 6,* 7–22. **Retrieved from http://journal.sjdm.org/11/rh13/rh13.pdf**

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

Gigerenzer, G., & Brighton, H. (2009). *Homo heuristicus*: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The Take the Best Heuristic. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75–95). New York, NY: Oxford University Press.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making, 6,* 100–121. **Retrieved from http://journal.sjdm.org/11/rh15/rh15.pdf**

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making, 6,* 23–42. **Retrieved from http://journal.sjdm.org/11/rh4/rh4.pdf**

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, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90. doi:10.1037/0033-295X.109.1.75

Hilbig, B. E. (2008). Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality, 42,* 1641–1645. doi:10.1016/j.jrp.2008.07.001

Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making, 5,* 272–284. **Retrieved from http://journal.sjdm.org/10/rh5/rh5.pdf**

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2010). One-reason decision making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 123–134. doi:10.1037/a0017518

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2012). A matter of time: Antecedents of one-reason decision making based on recognition. *Acta Psychologica, 141,* 9–16. doi:10.1016/j.actpsy.2012.05.006

Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition, 43,* 659–671. doi:10.3758/s13421-014-0493-z

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology, 55,* 394–401. doi:10.1027/1618-3169.55.6.394

Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- versus evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1296–1305. doi:10.1037/a0016565

Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making, 22,* 510–522. doi:10.1002/bdm.644

Hilbig, B. E., & Richter, T. (2011). *Homo heuristicus* outnumbered: Comment on Gigerenzer and Brighton (2009). *Topics in Cognitive Science, 3,* 187–196. doi:10.1111/j.1756-8765.2010.01123.x

Hilbig, B. E., Scholl, S. G., & Pohl, R. F. (2010). Think or blink—Is the recognition heuristic an "intuitive" strategy? *Judgment and Decision Making, 5,* 300–309. **Retrieved from http://journal.sjdm.org/10/rh6/rh6.pdf**

Horn, S. S., Pachur, T., & Mata, R. (2015). How does aging affect recognition-based inference? A hierarchical Bayesian modeling approach. *Acta Psychologica, 154,* 77–85. doi:10.1016/j.actpsy.2014.11.001

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40,* 1163–1177. doi:10.3758/s13421-012-0226-0

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review, 21,* 1272–1280. doi:10.3758/s13423-014-0608-3

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75,* 70–98. doi:10.1007/S11336-009-9141-0

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University Press.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28,* 3049–3067. doi:10.1002/sim.3680

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2010). From recognition to decisions: Extending and testing recognition-based models for multialternative inference. *Psychonomic Bulletin & Review, 17,* 287–309. doi:10.3758/PBR.17.3.287

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80,* 205–235. doi:10.1007/s11336-013-9374-9

Matzke, D., Lee, M. D., & Wagenmakers, E.-J. (2013). Multinomial processing trees. In M. D. Lee & E.-J. Wagenmakers (Eds.), *Bayesian cognitive modeling: A practical course* (pp. 187–195). New York, NY: Cambridge University Press.

McCloy, R. A., Beaman, C. P., Frosch, C. A., & Goddard, K. (2010). Fast and frugal framing effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1043–1052. doi:10.1037/a0019693

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Personality Psychological Review, 80,* 252–283. doi:10.1037/h0035002

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42,* 42–54. doi:10.3758/BRM.42.1.42

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19,* 333–346. doi:10.1002/bdm.531

Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 923–935. doi:10.1037/0278-7393.30.4.923

Odum, A. L. (2011). Delay discounting: Trait variable? *Behavioral Processes, 87,* 1–9. doi:10.1016/j.beproc.2011.02.007

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90,* B1–B9. doi:10.1016/S0010-0277(03)00141-0

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica, 125,* 99–116. doi:10.1016/j.actpsy.2006.07.002

Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making, 21,* 183–210. doi:10.1002/bdm.581

Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 983–1002. doi:10.1037/0278-7393.32.5.983

Pachur, T., Mata, R., & Schooler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging, 24,* 901–915. doi:10.1037/a0017211

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology, 2*(147), 1–14. doi:10.3389/fpsyg.2011.00147

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19,* 251–271. doi:10.1002/bdm.522

Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making, 6,* 423–438. **Retrieved from http://journal.sjdm.org/11/rh19/rh19.pdf**

Pohl, R. F., Erdfelder, E., Hilbig, B. E., Liebke, L., & Stahlberg, D. (2013). Effort reduction after self-control depletion: The role of cognitive resources in use of simple heuristics. *Journal of Cognitive Psychology, 25,* 267–276. doi:10.1080/20445911.2012.758101

Pohl, R. F., von Massow, F., & Beckmann, B. (2015). *Developmental differences in using an ecologically valid decision strategy: The case of the recognition heuristic*. Manuscript submitted for publication.

Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 150–162. doi:10.1037/0278-7393.32.1.150

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality, 43,* 137–145. doi:10.1016/j.jrp.2008.12.015

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., & Bröder, A. (2007). Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting, 23,* 415–426. doi:10.1016/j.ijforecast.2007.05.006

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review, 22,* 391–407. doi:10.3758/s13423-014-0684-4

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–628. doi:10.1037/0033-295X.112.3.610

Shiloh, S., Koren, S., & Zakay, D. (2001). Individual differences in compensatory decision-making style and need for closure as correlates of subjective decision complexity and difficulty. *Personality and Individual Differences, 30,* 699–710. doi:10.1016/S0191-8869(00)00073-8

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software, 12,* 1–16. **Retrieved from www.jstatsoft.org/v12/i03/paper**

Witkin, H. A., Goodenough, D. R., & Karp, S. A. (1967). Stability of cognitive style from childhood to young adulthood. *Journal of Personality and Social Psychology, 7,* 291–300. doi:10.1037/h0025070

Yechiam, E., & Busemeyer, J. R. (2008). Evaluating generalizability and parameter consistency in learning models. *Games and Economic Behaviour, 63,* 370–394. doi:10.1016/j.geb.2007.08.011