UPPSALA
UNIVERSITET

Additional Evidence for a Division of Labor in Single-cue Function Learning

Amos Pagin

Master's thesis in Psychology (15 credits)
Autumn term 2019
Department of Psychology
Uppsala University

Supervisor: Peter Juslin
Examiner: Anders Winman

Abstract

A novel division-of-labor hypothesis (DLH) suggests that single-cue function learning involves dynamic shifts between multiple, qualitatively distinct cognitive processes depending on task and judge properties (Juslin, Millroth, Sundh, & Nilsson, 2019). In this study, DLH was tested in an experiment where participants learned a quadratic function with either declarative or procedural stimulus- and response formats. Performance was measured again after a one-week forgetting interval. In agreement with DLH, participants seemingly solved the task using exemplar-based processes when the stimulus- and response modes were both declarative, and using rule-based or other processes when they were not. However, the forgetting interval did not induce the predicted shifts to rule-based processes.

Keywords: Function learning, multiple-cue judgment, fuzzy-trace theory

Many central problems in cognitive psychology can be characterized as problems of explaining, for some domain of cognition, how the cognitive system optimizes its performance in solving inductive problems common in that domain. In broad terms, inductive problems arise whenever the cognitive system seeks to infer the value of some variable but lacks sufficient information, or sufficient cognitive resources, to perform a deductively valid computation, forcing it instead to rely on weaker forms of inference in order to produce a suitable estimate (e.g., Evans & Stanovich, 2013; Holland, Holyoak, Nisbett, & Thagard, 1989). Because absence of task-relevant information is the norm rather than the exception, inductive problems abound in diverse areas of cognitive psychology, including causal cognition (Penn & Povinelli, 2007; Perales & Shanks, 2007; Sloman, 2005), categorization (Goldstone, Kersten, & Carvalho, 2012; Kruschke, 2005; Medin & Smith, 1984), language processing and language acquisition (Chater & Manning, 2006; Pinker, 1995), and judgment and decision-making (Brehmer, 1974; Gigerenzer & Brighton, 2009; Kahneman & Frederick, 2005; Tversky & Kahneman, 1992). Over the past decades, a number of studies on *function learning* have shed light on how the cognitive system optimizes performance on a particular class of inductive problems, namely those that involve predicting a continuous response variable on the basis of a known continuous stimulus variable. To this end, several models have been proposed attempting to capture the cognitive processes involved in function learning, emphasizing either rule-based processes (e.g., Carroll, 1963; Brehmer, 1974), exemplar-based processes (e.g., Busemeyer, Byun, Delosh, & McDaniel, 1997), or hybrid processes involving both rule-based and exemplar-based representations (Delosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004).

In this thesis, a novel hypothesis regarding human function learning will be tested, and some of its implications will be explored. The hypothesis, first presented in Juslin, Millroth, Sundh, and Nilsson (2019), suggests that function learning is not reducible to a single (hybrid or non-hybrid) process as assumed in previous models, but that it involves a division of labor between multiple qualitatively distinct cognitive processes, resulting in dynamic shifts between

rule-based and exemplar-based processes depending on both task-specific and judge-specific properties. Because rule-based and exemplar-based processes involve different computations, it is predicted that different judgment patterns will emerge depending on which process is engaged at the time of judgment. Accordingly, the hypothesis predicts data that are difficult to reconcile with established function learning models that assume only a single process. In the present study, results from an experiment will be presented displaying data that correspond poorly to currently established models of function learning, supporting instead the division-of-labor hypothesis.

Single-Cue Function Learning

Function learning forms the basis of many routine estimation tasks encountered in everyday life. Such tasks might include, for instance, predicting the braking distance of a car based on its current speed, the weight of a suitcase based on its perceived heaviness when held, or the temperature of the water coming out of a faucet based on the setting on the control valve (e.g., Busemeyer et al., 1997; Kalish et al., 2004). Indeed, learning functional relations between variables is all but crucial in understanding and predicting the world, and it is difficult to imagine what life would be like if the ability to do so was suddenly lost. Given an adequate amount of learning trials, predictions of this sort can often be carried out almost effortlessly, and often with a high degree of accuracy, even for magnitudes of the stimulus dimension not previously encountered (Kalish et al., 2004). Furthermore, everyday experience indicates that our ability to learn functional relationships is not limited to functions of a certain form (e.g., positive linear relations), but that it accommodates a wide variety of functions, ranging from positive and negative linear relations to non-linear, non-monotonic, and non-continuous functions (see Busemeyer et al., 1997 for a review). For researchers in cognitive psychology, these considerations raise a number of questions. For instance, when a person learns to predict continuous response variables in this manner, what is it that has been learned, and how? What sort of representations form the basis for function learning, and what computations are performed on these representations? Is function learning governed by a single cognitive process, or by several? And if the answer is ”several”, what are the factors that reinforce shifts between different cognitive processes?

In order to study human performance in single-cue function learning, researchers have devised a number of function learning tasks where participants must learn to predict a continuous response variable (the criterion) on the basis of a single continuous stimulus variable (the cue). A standard approach has participants first read a cover story, and then complete a training phase consisting of a discrete set of learning trials. On each learning trial, a cue value (i.e., a magnitude of the stimulus dimension) is presented, and the participants must provide a point estimate of the response magnitude. Corrective feedback is presented after each trial, and participants typically learn to improve the accuracy of their point estimates on the basis of this feedback. Once the training phase is completed, participants complete a test phase in which they again provide point estimates of the response magnitude, but with no corrective feedback. The test phase might also

contain novel cue values not previously trained on, so that interpolation or extrapolation is necessary in order to produce a judgment. Performance is then assessed by measuring the deviations between judgments and criterion values in the test phase.

Using this paradigm and similar variations, researchers have established a number of robust findings with regard to function learning performance (see Busemeyer et al., 1997; Kalish et al., 2004; Lucas, Griffiths, Williams, & Kalish, 2015 for reviews and discussion). With regard to overall performance, systematic continuous functions are easier to learn than arbitrary cue-criterion associations (Busemeyer et al., 1997), linear functions are easier to learn than non-linear functions (Brehmer, 1974; Brehmer, Alm, & Warg, 1985; Byun, 1996), increasing linear functions are easier to learn than decreasing linear functions (Brehmer, 1971, 1976), and montone functions tend be more easily learned than non-monotone functions (Lucas et al., 2015). Cyclic functions are apparently particularly difficult to learn (Bott & Heit, 2004; Byun, 1996; Kalish, 2013). Results further indicate that interpolation performance for novel cue values within the training range is nearly as accurate as performance on trained cue values; however, extrapolation performance beyond the training range is generally less accurate (Busemeyer et al., 1997; Lucas et al., 2015).

Cognitive Processes in Single-Cue Function Learning

Early research on function learning suggested that the learning process involves abstraction of explicit rules that summarize the trained cue-criterion pairings in a manner analogous to statistical regression (e.g., Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991). Initially, a compelling motivation for the rule-based approach was that it provided a plausible explanation for people's ability to extrapolate beyond training data. An early proposal in this vein was outlined by Carroll (1963), who characterized function learning as a form of parameter estimation in a polynomial regression model; this proposal was later expanded upon by Brehmer (1974), who suggested that people share a common implicit hierarchy of hypotheses with regard to the shape of the function they are learning. This expanded model, sometimes referred to as the *polynomial hypothesis-testing model* (e.g., Delosh et al., 1997; McDaniel, Dimperio, Griego, & Busemeyer, 2009), was able to provide an account of people's performance in function learning tasks, while at the same time explaining why certain function forms prove more difficult to learn than others.

However, critics have emphasized a number of limitations with the rule-abstraction approach. A first objection is that the rule-based approach does not adequately capture extrapolation performance. For instance, Delosh et al. (1997) measured extrapolation performance for both linear, quadratic, and exponential functions, with results indicating that polynomial models were not accurate in describing extrapolation performance for any of the function forms investigated. Secondly, it has been argued that rule-based models provide little insight with regard to the learning process itself. The polynomial hypothesis-testing model does not specify, on a trial-by-trial basis, how hypotheses about function shape are selected or rejected, or how parameters estimates are adjusted in the light of corrective feedback. A third point of

criticism is that the inherent similarities between function learning and category learning suggest that a common theoretical framework should be able to explain both phenomena, and that theories of function learning should thus be founded on the better established principles of category learning (Busemeyer et al., 1997).

Objections of this nature spurred the development of ALM (Associative-Learning Model; Busemeyer et al., 1997), a connectionist model of function learning based not on rule abstraction, but on principles of associative learning borrowed from an earlier model, ALCOVE (Attention Learning Covering Map; Kruschke, 1992), which was developed to model category learning. In ALM, function learning is construed as a process in which direct associations between stimulus and response magnitudes are stored in memory in the form of exemplars. Accordingly, the process of producing a value of the response dimension given a cue value is not mediated by rule-based processes, but involves the retrieval of the associated response magnitude from memory. As demonstrated in Busemeyer et al. (1997), ALM can successfully account for a number of findings in function learning, including the order of learning difficulty for different function forms. However, a significant limitation of ALM is that it contains no viable mechanism for extrapolation, and hence cannot adequately account for extrapolation performance (Busemeyer et al., 1997, Delosh et al., 1997). This limitation instigated the development of hybrid models, in which function learning is conceptualized as involving both rule-based and instance-based processes. To date, two hybrid models have been proposed, with both being able to account for an impressive amount of function-learning data: EXAM (EXtrapolation Associative Model; Delosh et al., 1997) and POLE (Population Of Linear Experts; Kalish et al., 2004).

EXAM is an adaptation of ALM, in which the associative learning mechanism has been supplemented with a response rule producing linear extrapolations (Delosh et al., 1997). The response rule, which was motivated by Wagenaar and Sagaria's (1975) observation that extrapolation tends to be approximately linear even for nonlinear functions, works by matching novel cues from outside the training range with the closest neighboring training cues. Predictions from those training cues are then retrieved, and a line is estimated from these predictions. Accordingly, the response to a novel cue is produced by retrieving the relevant value on this estimated line. By contrast, in POLE, trained cue values are not associated with their corresponding criterion values, but with an array of linear candidate functions referred to as "experts", each one predicting the target magnitude (Kalish et al., 2004). On any given trial, all experts compute their respective outputs, but only one expert's output is selected as the response. The probability of an expert's output being selected is determined by two factors: First, each expert has a stimulus-independent probability of being selected, which is taken to embody a combination of a priori expectations and learned preferences. Second, POLE assumes a weighted connection between each expert and representations of previously encountered stimuli. These two factors jointly determine a strength value for each expert, and for any given trial, the output of the expert with the highest strength value is selected as the response. When corrective feedback is made available, learning occurs by adjusting both the stimulus-independent biases

and the stimulus-specific connection weights in the light of the observed error. Additionally, a noteworthy characteristic of POLE is that the same stimulus magnitude can be associated with different experts, especially if the stimulus is presented in different contexts. This characteristic embodies an important concept often referred to as *knowledge partitioning*—the notion that knowledge can be separated into independent parcels that may contain contradictory information. The ability to account for knowledge partitioning is an attractive aspect of POLE, as there is evidence to suggest that knowledge partitioning occurs in several domains, including category learning (Lewandowsky & Kirsner, 2000; Yang & Lewandowsky, 2004), expert judgment (Lewandowsky et al., 2000), and function concept acquisition (Lewandowsky, Kalish, & Ngang, 2002).

Evidence for a Division of Labor

Juslin et al. (2019) argue that existing models of function learning share a common assumption: They conceive of function learning as being governed by a single process. Even hybrid models such as EXAM and POLE, which involve both rule-based and exemplar-based representations, are founded on the assumption that a single hybrid process is applied to all single-cue function-learning tasks. As an alternative hypothesis, Juslin et al. (2019) suggest that function learning does not rely on a single process, but on an adaptive division of labor between multiple, qualitatively distinct cognitive processes, and that the cognitive system shifts between different processes as required in order to optimize function learning performance.

Justification for this division-of-labor hypothesis comes from research on multiple-cue judgment, where participants are required to map multiple cues, which can be either categorical or continuous, to a continuous criterion variable. In this research, evidence has accumulated that participants shift dynamically between rule-based and exemplar-based processes depending on the properties of the task at hand (Envkist, Newell, Juslin, & Olsson, 2006; Helversen & Rieskamp, 2008; Hoffman, Helversen, & Rieskamp, 2014; Juslin, Karlsson, & Olsson, 2008; Juslin, Olsson, & Olsson, 2003; Karlsson, Juslin, & Olsson, 2007; Pachur & Olsson, 2012; Platzer & Bröder, 2013), and properties of the judge (Hoffman et al., 2014; Little & McDaniel, 2015). Although work remains to be done in modeling the real-time learning process in which such representational shifts are instigated, it is thought that the rule-based processes in controlled judgment are largely constrained to linear additive combinations of cues (see Juslin et al., 2008 for a discussion), and that shifts to other cognitive resources, such as exemplar memory (Helversen & Rieskamp, 2008; Juslin et al., 2008; Karlsson et al., 2007; Olsson, Juslin, & Olsson, 2006; Pachur & Olsson, 2012) and pattern matching (Helversen & Rieskamp, 2008; Kelley & Busemeyer, 2008), occurs as the cognitive system recognizes that mastering the task at hand involves learning relations that are nonlinear, or cue combinations that are non-additive.

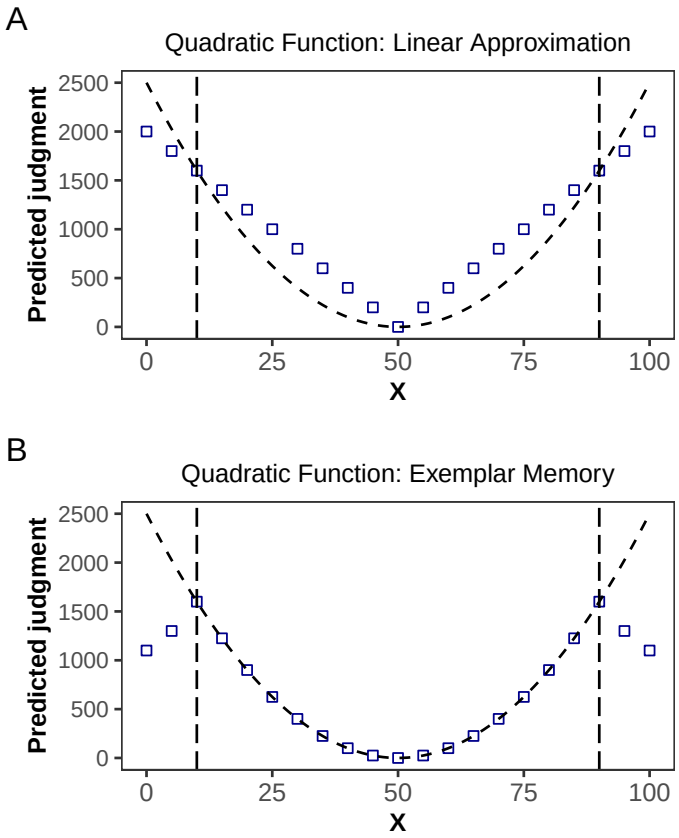

Figure 1. Predictions by a piece-wise linear rule-based model (Panel A) and an exemplar-based model (Panel B) in a single-cue function learning task, where participants use an X-value to predict a Y-value. The participants train to perform the task in the interval of $X[10,90]$. In a later test phase, they are required to extrapolate beyond this region. The piece-wise linear rule-based model (Panel A) predicts a linear distortion of the function, whereas the exemplar-based model (Panel B) predicts an inability to extrapolate.

Given this background, Juslin et al. (2019) propose that a generalization of the results from research on multiple-cue judgment implies that similar shifts between rule-based and exemplar-based processes occurs in single-cue function learning, implying both between-task heterogeneity (participants may address different function-learning tasks with qualitatively distinct cognitive processes), within-task heterogenetiy (participants may address different regions of the same function with qualitatively distinct cognitive processes), and within-belief heterogeneity (participants may respond to questions about the underlying function in ways implying contradictory beliefs, depending on what processes are engaged by the question format). Additionally, because the rule-based and exemplar-based processes are known to invoke different computations, the judgment patterns resulting from each process should be a priori predictable (Juslin et al., 2019). Following this line of thought, Figure 1 illustrates predictions from Juslin et al. (2019) with respect to human judgment patterns for a function-learning task involving a quadratic function. The illustrations assume that the participants train with corrective feedback to predict a criterion (y) from a single cue (x), with X ranging from 10 to 90; in a later

test phase, participants predict y for the entire X-range between 0 and 100 without feedback, thus necessitating extrapolation for trials where $X < 10$ or $X > 90$.

As seen in Figure 1A, an emphasis on rule-based processes should reveal characteristic linear distortions, as well as an ability to extrapolate in the direction of the function. By contrast, an emphasis on exemplar-based processes should result in judgments that capture the function with great fidelity within the training range but produce poor extrapolation performance with characteristic drops in the predictions for the cue values outside the training range, as illustrated in Figure 1B.

In order to test the division-of-labor hypothesis, Juslin et al. (2019) devised two versions of a function-learning task, where one version of the task was designed to induce mainly rule-based processes, and the other version was designed to induce mainly exemplar-based processes. This was accomplished by drawing from research on category learning, where studies have indicated that *stimulus confusion*—the inability to distinguish between stimuli in an absolute identification task—is a variable that predicts whether participants engages in rule-based versus exemplar-based processes in category learning: When stimuli are confusable, participants favor rule-based processes, whereas they favor exemplar-based processes when each stimulus is clear, distinct, and easily identifiable (Rouder & Ratcliff, 2004). Adapting stimulus confusion to a function-learning task, one version of the task utilized procedural stimulus and response formats, where cue values were indicated by a mark on a horizontal bar (corresponding to a percentage of the maximum length of the bar between 0% and 100%), and participants indicated their point estimates were by adjusting the lightness of a square area from white (minimum) to black (maximum) by clicking on a bar with all shades in between white and black. The second version of the task utilized declarative stimulus and response formats, where both cue values and point estimates were expressed in terms of numerals.

Results were largely in line with the predictions: When the participants learned a quadratic (U-shaped) function, their judgment patterns tended towards those in Figure 1A for the procedural format, but towards those in Figure 1B for the declarative format, thus supporting the notion of between-task process heterogeneity in function learning. Additionally, when the participants learned an inverse quadratic function, results for the declarative format indicated that although participants were unable to extrapolate for high x-values, they were able to extrapolate for low x-values, in a region of the function where y could be construed as an approximately positive linear function of x. This is turn suggests that participants find it easier to engage the abstraction of linear rules in regions where the function is increasing, and that participants may shift between exemplar-based and rule-based processes depending on the region of the function they are learning, hence supporting the notion of within-task process heterogeneity.

However, with regard to the results obtained in Juslin et al. (2019), two questions remain unresolved. The first question concerns the robustness of the results: Will a similarly constructed experiment, designed to induce either exemplar-based or rule-based processes, yield similar results? The second question concerns the moderating factors for shifts between exemplar- and rule-based processes. In Juslin et al. (2019), the function learning tasks involved either

declarative or procedural stimulus- and response modes, but it was not explored how function learning performance is affected when the stimulus- and response modes are crossed (i.e., using a declarative stimulus mode and a procedural response mode, or vice versa). Both of these questions will be addressed in the present study.

An additional hypothesis, derived from fuzzy-trace theory, suggests that reliance on rule-based processes should increase over time. According to fuzzy-trace theory, people form two types of mental representations of past events: Verbatim traces and gist traces (e.g., Reyna & Brainerd, 1995; Brainerd & Reyna, 1993). Whereas verbatim traces are precise, detailed representations of past events, gist traces are taken to be vague, fuzzy and imprecise, implying that the overall gist is remembered but precise information about the event is not. While verbatim traces may fade quickly from memory (either because verbatim traces degrade over time, or because verbatim knowledge is not properly encoded in the first place), gist traces are more persistent, and can remain in memory for extended time periods (Reyna & Brainerd, 1995). An implication for function learning is that, because exemplar-based processes in function learning rely on verbatim traces (i.e., precise, detailed recollections of exemplars), the degradation of verbatim traces over time should instigate a shift from exemplar-based processes to rule-based processes if the same function learning task is repeated following a forgetting interval. It is thus hypothesized that participants who learn a quadratic function using exemplar-based processes will display judgment patterns similar to those in Figure 1A when first completing the function learning task, but that their judgment patterns will resemble those in Figure 1B if function learning task is completed again at a later point in time, after the verbatim traces have degraded.

Aims of the Present Study

In sum, the present study is a follow-up to Juslin et al. (2019) with three distinct aims. A first aim is to replicate results in Juslin et al. (2019) that suggest between-task process heterogeneity. A second aim is to explore moderating factors for shifts between exemplar-based and rule-based processes by crossing the stimulus- and response modes utilized in Juslin et al. (2019) (i.e., using a declarative stimulus mode and a procedural response mode, and vice versa). A third aim is to test the hypothesis, derived from fuzzy-trace theory, that participants who complete a function-learning task utilizing primarily exemplar-based processes will shift to utilizing mainly rule-based processes if the task is completed again following a forgetting interval. The shift is predicted from the degradation of verbatim traces, which form the basis for exemplar-based judgments.

To accomplish this, an experiment is conducted based around a function-learning task involving a quadratic function. It is predicted that participants who complete the task using declarative stimulus- and response formats will display judgment patterns that capture the function with great fidelity within the training range, but display extrapolation judgments negatively correlated with the true function form, similar to those in Figure 1B, whereas participants who complete the task using procedural stimulus- and response formats will display

judgment patterns using piecewise linear approximations resulting in linear distortions similar to those depicted in Figure 1A. For participants who complete the task using mixed formats (e.g., a declarative stimulus mode and a procedural response mode), no predictions are made, as this part of the study is mainly exploratory. Finally, it is predicted that linear distortions, similar to those in Figure 1A, will be present for all participants when the function-learning task is completed again after a forgetting interval, due to an increased reliance on rule-based processes.

Method

Participants

Seventy-four voluntary participants, 52 female and 22 male, with an average age of 30 years ($SD = 12.3$ years) received either a cinema ticket (worth approximately 15 US dollars) or course credit as compensation for their participation.

Design

The experiment involved a 2x2x2 mixed factorial design, with Stimulus Mode (declarative vs. procedural) and Response Mode (declarative vs. procedural) as between-subjects variables, and Test Session (first session vs. second session) as independent within-subjects variable. The dependent measures consisted of the point estimates provided by the participants, as well as the accuracy of said estimates in terms of the Root Mean Square Deviation (RMSD) between judgments and criteria. Participants were randomized to one of the four between-subject cells.

Materials and Procedure

Each participant partook in two separate test sessions, with the second session taking place after a one-week forgetting interval. In botH-test sessions, participants completed a predetermined number of trials in which the task was to provide point estimates of a criterion (y) given a cue value (x) that varied between trials. Formally, the relationship between x and y was given by the equation $y = (50 - x)^2$, denoting a quadratic (U-shaped) function. The same function was used on all trials, and neither the shape of the function nor its equation were disclosed to the participants.

In the first test session, participants read a cover story stating that their task is to learn the relationship between the concentration of a substance called *Strontium* (x) in the blood stream, and the severity of a disease called *Brunswik's agony* (y). Participants then completed a training phase with corrective feedback consisting of 102 trials. In each trial, participants were asked the question: "When the value of Strontium is x, what is the value of Brunswik's agony?" Values of X were drawn randomly from the range $[10, 90]$ in intervals of five (i.e., 10, 15, 20, etc.), and each such value was presented six times, each time in a separate trial. The cue values were not partitioned into separate blocks, but were presented in a single randomized sequence. Corrective feedback was given after each trial by displaying the criterion value.

Because stimulus mode and response mode were experimentally manipulated, both the manner in which the cue value was presented and the manner in which participants indicated their point estimates varied between the experimental conditions. For the participants who were assigned a declarative stimulus mode, X was indicated using numerals (e.g., "When the value of Strontium is 45, what is the value of Brunswik's agony?"), and for the participants who were assigned a procedural stimulus mode, X was indicated using a mark on a horizontal bar (corresponding to a percentage of the maximum length of the bar between 0% and 100%). With regard to response mode, participants who were assigned a numerical response mode keyed in their point estimates using numerals, and participants who were assigned a procedural response mode indicated their point estimate by adjusting the lightness of the shade of a square area from white (minimum) to black (maximum), by clicking on a bar representing all shades in between white and black. See Figures S1 and S2 in the Appendix for screenshots from the desktop application used for the experiment depicting the differences between declarative and procedural stimulus and response formats.

After the training phase, the participants completed a test phase which was identical to the training phase, except that the test phase contained no corrective feedback, and the range of X was expanded from [10,90] to [0,100], thus meaning that participants were required to extrapolate when confronted with cue values not trained on. Additionally, each item was presented on only two trials, instead of on six trials, thus resulting in a total of 42 trials in the test phase. Finally, the second test session, occurring approximately one week later, was identical to the first test session, except that the second session contained no training phase.

Results

Performance was analyzed in terms of Root Mean Square Deviations (RMSDs) between judgments and criteria. Lower RMSD values indicate better performance (i.e., smaller differences between judgments and criteria) than higher RMSD values. Because the RMSD distributions failed to satisfy the relevant statistical assumptions (e.g., normality, homogeneity of variance) in almost all cases, non-parametric tests were conducted instead of a ANOVAs. Aside from statistical tests, the observed judgments were also fitted to a priori predictions from three different cognitive models—one model assuming that that participants correctly extracted the underlying function, one model assuming that participants relied on piecewise linear approximations, and one model assuming that participants relied on exemplar memory. Performance results are presented first, followed by modeling results (with specifications of the models) in a separate section.

Performance at the first test session

With regard to the first test session, performance was similar between conditions when comparing the point estimates for cue values within the training range, and the participants learned to approximate the underlying function irrespective of which stimulus- and response

mode they had been assigned. As indicated in Figure 2A, the participants in the declarative–declarative condition (the first term refers to the stimulus mode, and the second term to the response mode) were not able to extrapolate in the direction of the function, displaying instead extrapolation judgments negatively correlated with the function form, as predicted by the division-of-labor hypothesis (see Figure 1B). However, it was also predicted that the judgment pattern for the procedural–procedural condition would reveal linear distortions, as illustrated in Figure 1A. Contrary to this prediction, Figure 2D does not suggest the presence of linear distortions, indicating instead that participants successfully extracted the underlying function.

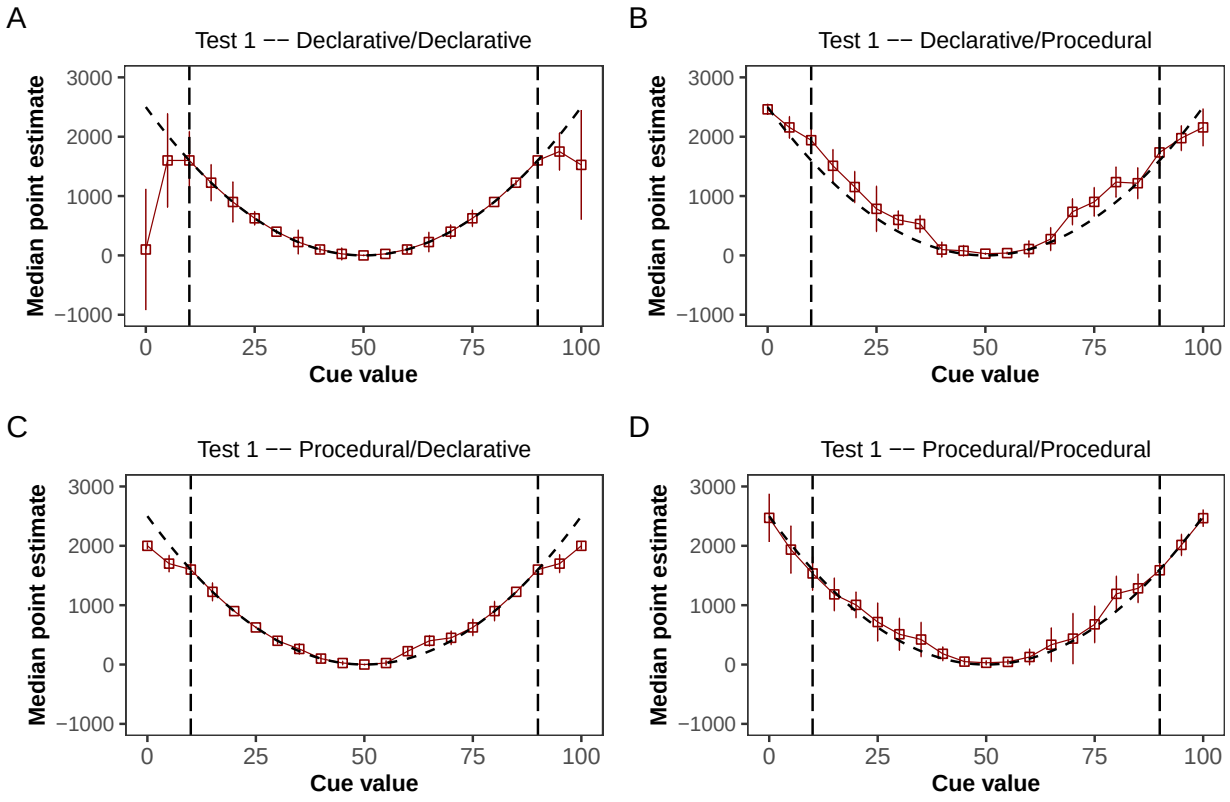

Figure 2. Median point estimates for each cue value in each of the cells in the 2 x 2 factorial design, on the first test session. The first term in the panel heading identifies the stimulus format, and the second term identifies the response format. Error bars indicate interquartile ranges (IQRs). The dashed U-shape represents the underlying function form, and the vertical lines represent boundaries for the training region.

An initial question is whether or not overall performance at the first test session was significantly affected by the choice of stimulus mode or by the choice of response mode. A Mann-Whitney U-test with the independent variable Stimulus Mode indicated that the RMSD was significantly smaller for the procedural stimulus mode ($Mdn = 259.86$) than for the declarative stimulus mode ($Mdn = 534.65$), $U = 877, p = .004, r = .34$), thus indicating that performance at the first test session was better with the procedural stimulus mode than with the declarative stimulus mode. By contrast, a Mann-Whitney U-test with the independent variable Response Mode indicated no significant difference in RMSD between the declarative response format ($Mdn = 260.48$) and the procedural response format ($Mdn = 413.88$), $U = 513, p = .191$.

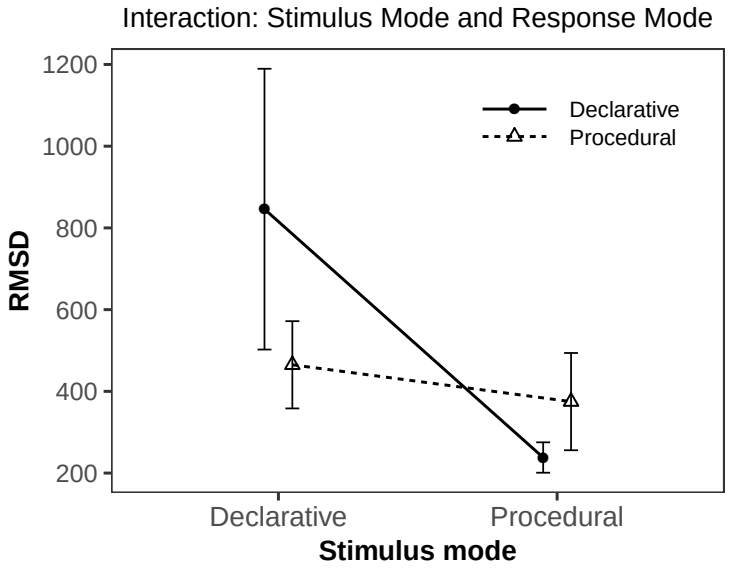

Figure 3. Median Root Mean Square Deviations (RMSDs) for each of the cells in the 2 x 2 factorial between-subjects design, with error bars representing Interquartile Ranges (IQRs).

A second question is whether or not there were significant performance differences between each cell of the 2x2 factorial design. Figure 3 plots median RMSDs for each cell and suggests that performance in the declarative-declarative condition was worse than performance in any other cell of the design. A Kruskal-Wallis H-test was conducted on the RMSD between judgments and criterion, where the two independent between-subjects variables Stimulus Mode (declarative vs. procedural) and Response Mode (declarative vs. procedural) were collapsed into the single independent variable Experimental Condition with four levels. The test revealed a significant performance difference between the cells ($\chi^2(3) = 11.57, p = .009, \eta^2 = .128$). Post-hoc tests were conducted in terms of Bonferroni-corrected pairwise Mann-Whitney U-tests. The post-hoc tests yielded only one significant result, indicating that the RMSD for procedural–declarative condition was significantly lower ($Mdn = 238$) than the RMSD for the declarative–procedural condition ($Mdn = 465$), $U = 28, p < .001, r = .68$. See table S1 in the Appendix for a complete table of post-hoc tests relating to the first session.

As seen in Figure 2, the performance differences between the cells are most clearly pronounced in the extrapolation regions. This raises the question of whether the significant findings are driven primarily by differences in extrapolation performance, or whether performance differences remain statistically significant even when the training region is examined in isolation (i.e., when excluding all extrapolation items). In order to answer this question, additional analyses restricted to the training region were performed. To this end, a Mann-Whitney U-test with the independent variable Stimulus Mode indicated no significant difference in RMSD between the declarative stimulus mode ($Mdn = 368$) and the procedural stimulus mode ($Mdn = 513$), $U = 781, p = .081$). By contrast, a Mann-Whitney U-test with the independent variable Response Mode indicated that the RMSD was significantly smaller for the

declarative response mode ($Mdn = 197$) than for the procedural response mode ($Mdn = 305$), $U = 265, p < .001$, thus indicating that performance on trained cue items was better with the declarative response mode than with the procedural response mode. A Kruskal-Wallis H-test, where the two individual variables Stimulus Mode and Response Mode were collapsed into the single independent variable Experimental Condition with four levels, revealed a significant performance difference between the conditions ($\chi^2(3) = 21.53, p < .001, \eta^2 = .277$). Post-hoc tests were conducted in terms of Bonferroni-corrected pairwise Mann-Whitney U-tests, and revealed that performance was significantly better in the procedural–procedural condition ($Mdn = 305$) compared to the declarative–procedural condition ($Mdn = 405$), $U = 36, p < .001, r = .674$. It was also better in the procedural–declarative condition ($Mdn = 150$) compared with the declarative–procedural condition ($Mdn = 405$), $U = 270, p < .001, r = .758$. See table S2 in the Appendix for a complete table of post-hoc tests.

In sum, the analysis of performance on the first test session indicated that the procedural stimulus mode resulted in better performance than the declarative stimulus mode. However, this effect was primarily driven by the comparatively low extrapolation performance in the declarative–declarative condition. Consequently, when extrapolation items were excluded from the analysis, the difference in performance between the two stimulus modes was no longer significant. Additionally, the analysis revealed no significant difference in performance between the two response modes. However, when extrapolation items were excluded from the analysis, the declarative response mode yielded significantly better performance than the procedural response mode.

Performance at the second test session

With regard to the second test session, performance was again similar between conditions, and indicated that participants had retained the ability to approximate the underlying function even after the one-week forgetting interval. As indicated in Figure 4A, the declarative-declarative condition again demonstrated extrapolation judgments that did not correspond well with the underlying function form. An interesting result is that the judgment pattern for the procedural-procedural condition revealed linear distortions (Figure 4D). The reader is reminded that this judgment pattern was predicted for the first test session, but that the data—contrary to the prediction—-instead suggested that participants had successfully extracted the underlying function (Figure 2D).

As for the first test session, an initial question is whether or not overall performance was significantly affected by the choice of stimulus mode or by the choice of response mode. A Mann-Whitney U-test with the independent variable Stimulus Mode indicated no significant difference in RMSD between the declarative stimulus format ($Mdn = 513$) and the procedural stimulus format ($Mdn = 305$), $U = 713, p = .337$. By contrast, a Mann-Whitney U-test with the independent variable Response Mode indicated that the RMSD was significantly lower for the declarative response mode ($Mdn = 240$) compared with the procedural response mode

($Mdn = 538$), $U = 310, p < .001, r = .434$. Hence, performance on the second test session was thus significantly better with the declarative response mode compared to the procedural response mode.

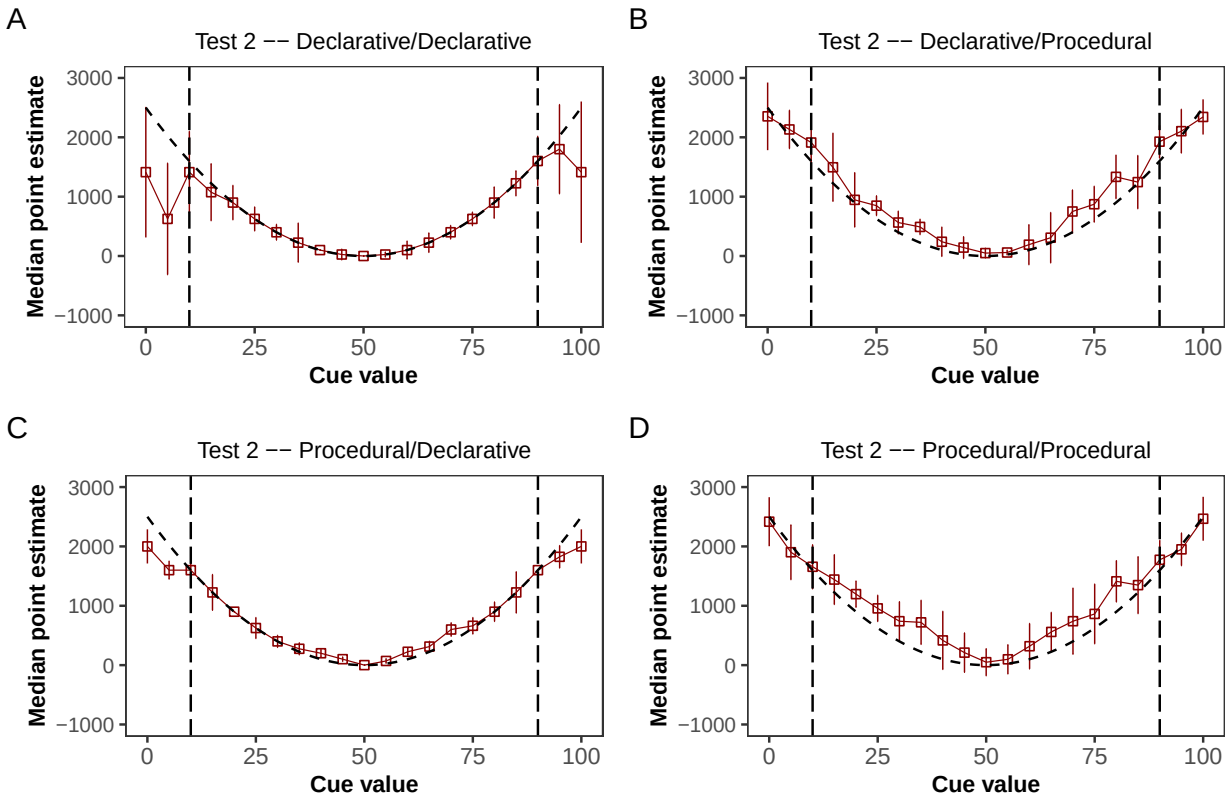

Figure 4. Median point estimates for each cue value in each of the cells in the 2 x 2 factorial design, for the second test session. The first term in the panel heading identifies the stimulus format, and the second term identifies the response format. Error bars indicate interquartile ranges (IQRs). The dashed U-shape represents the underlying function form, and the vertical lines represent boundaries for the training region.

A second question is whether or not there were significant performance differences between each cell of the 2x2 factorial design. Figure 5 plots median RMSDs for each cell, and suggests similar performance data as for the first test session, with the combination of a declarative stimulus mode and a declarative response mode yielding poorer performance than any other combination. A Kruskal-Wallis H-test was conducted on the RMSD between judgments and criteria, where the two independent between-subjects variables Stimulus Mode (declarative vs. procedural) and Response Mode (declarative vs. procedural) were collapsed into the single independent variable Experimental Condition with four levels. The test revealed a significant difference in RMSD between the conditions ($\chi^2(3) = 10.7, p = .014, \eta^2 = .115$). Bonferroni-corrected pairwise Mann-Whitney U-tests revealed that the procedural–declarative condition ($Mdn = 203$) performed significantly better than the declarative–procedural ($Mdn = 538$) condition, $U = 243, p < .001, r = .6$. No other post-hoc tests were found to be significant. Table S3 in the Appendix contains the complete table of post-hoc tests relating to the second test session.

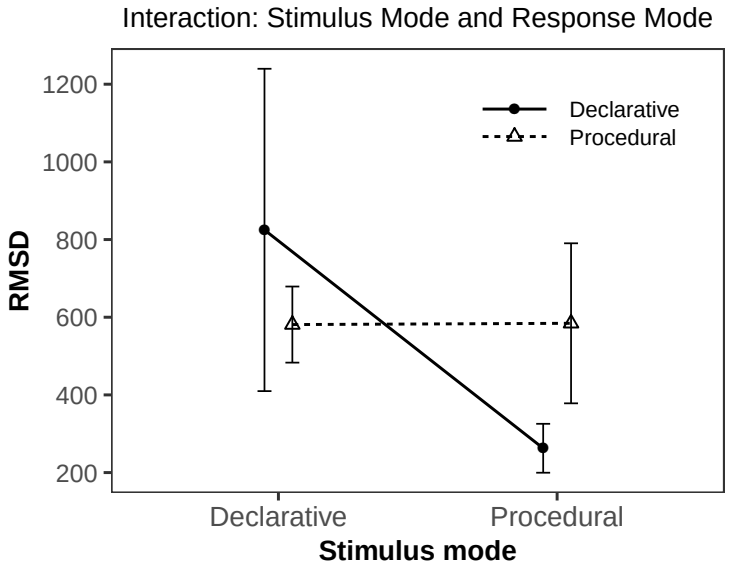

Figure 5. Median Root Mean Square Deviations (RMSDs) for each of the cells in the 2 x 2 factorial between-subjects design, with error bars representing Interquartile Ranges (IQRs).

Similar to the first test session, Figure 4 indicates that the performance differences between the cells were most clearly pronounced in the extrapolation regions. Hence, this raises again the question of whether the significant findings are driven mainly by differences in extrapolation performance, or whether there are significant differences in performance for trained cue values as well. A Mann-Whitney U-test on the RMSD between judgments and criteria with the independent variable Stimulus Mode indicated no significant difference between the declarative stimulus mode ($Mdn = 513$) and the procedural stimulus mode ($Mdn = 305$); however, a Mann-Whitney U-test on the RMSD between judgments and criteria with the independent variable Response Mode indicated that the RMSD was significantly lower with the declarative response mode ($Mdn = 240$) compared with the procedural response mode ($Mdn = 538$), $U = 310, p < .001, r = .434$. A Kruskal-Wallis H-test, where the two individual variables Stimulus Mode and Response Mode were again collapsed into the single independent variable Experimental Condition with four levels, revealed a significant performance difference between the four cells ($\chi^2(3) = 16.67, p < .001, \eta^2 = .21$). Bonferroni-corrected post-hoc tests revealed that the procedural–declarative condition ($Mdn = 203$) performed significantly better than the declarative–procedural condition ($Mdn = 538$), $U = 247, p < .001, r = .622$, and also better than the procedural–procedural condition ($Mdn = 490$), $U = 49, p < .001, r = .609$. See table S4 in the Appendix for a complete table of post-hoc tests.

In sum, the analysis of performance on the second test session did not reveal a significant difference in performance between the two stimulus modes (irrespective of whether extrapolation items were excluded or not). However, performance was better with the declarative response format than with the procedural response format, and this effect remained significant when extrapolation items were excluded.

Performance differences between the two test sessions

An additional question is how performance at the first test session (before the forgetting interval) compared to performance on the second test session (after the forgetting interval). Figure 6 indicates that performance was lower (i.e., RMSD was higher) after the forgetting interval in all experimental conditions, except for the declarative-declarative condition, where a small increase in performance was observed.

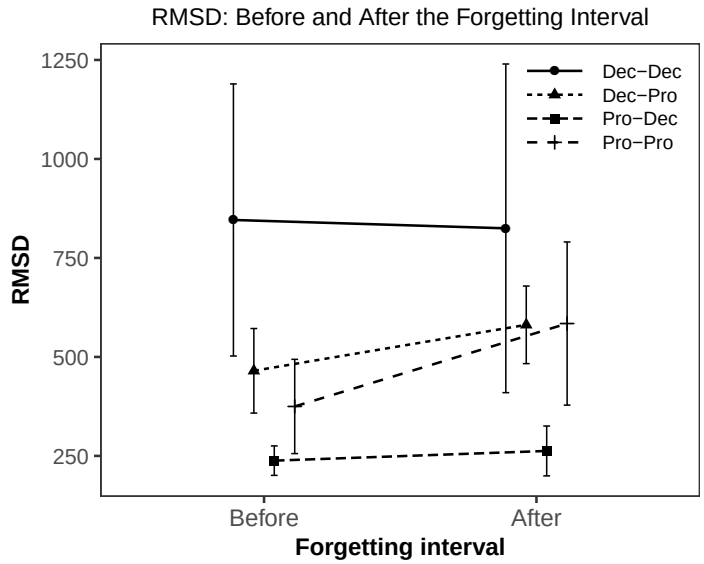

Figure 6. Median Root Mean Square Deviations (RMSDs) before and after the forgetting interval, for each of the cells in the 2 x 2 factorial between-subjects design. The first term in the legend headings identifies the stimulus mode, and the second term identifies the response mode. Error bars represent Interquartile Ranges (IQRs).

Wilcoxon signed rank test on the median RMSD between judgments and criteria, with the dependent variable Test Session (first session vs. second session), was conducted separately for each cell of the 2 x 2 factorial design. The tests revealed that the procedural–declarative condition performed significantly better before the forgetting interval ($Mdn = 238$) than after the forgetting interval ($Mdn = 263$), $W = 22, p = .002, r = .674$, and likewise for the procedural–procedural condition ($Mdn = 375$ before, $Mdn = 584$ after), $W = 18, r = .693$. For all other conditions, $p > .13$. See table S5 in the Appendix for a complete table of Wilcoxon signed-rank tests.

Additionally, a Kruskal-Wallis H-test, where the two individual variables Stimulus Mode and Response Mode were collapsed into the single independent variable Experimental Condition indicated no significant differences between conditions with regard to the RMSD differences between test sessions ($\chi^2(3) = 5.9, p = .117$).

Differences between trained cue values and exploration items

As mentioned, Figures 2 and 4 indicated that extrapolation performance may differ from performance on trained cue values. Moreover, such differences seem to be more pronounced in

some cells of the design than in others, with the combination of a declarative stimulus mode and a declarative response mode producing substantially poorer extrapolation performance than other combinations. Hence, one question is whether the differences in performance between exploration items and trained cue values is significantly affected by the choice of stimulus mode, or by the choice of response mode. In order to compare extrapolation performance with performance on trained cue values, the RMSDs for extrapolation items were subtracted from the RMSDs for training items, and analysis was then conducted on the resulting differences. As previously, the data failed to satisfy the relevant statistical assumptions (e.g., normality, homogeneity of variance), and thus non-parametric tests were conducted instead of ANOVAs.

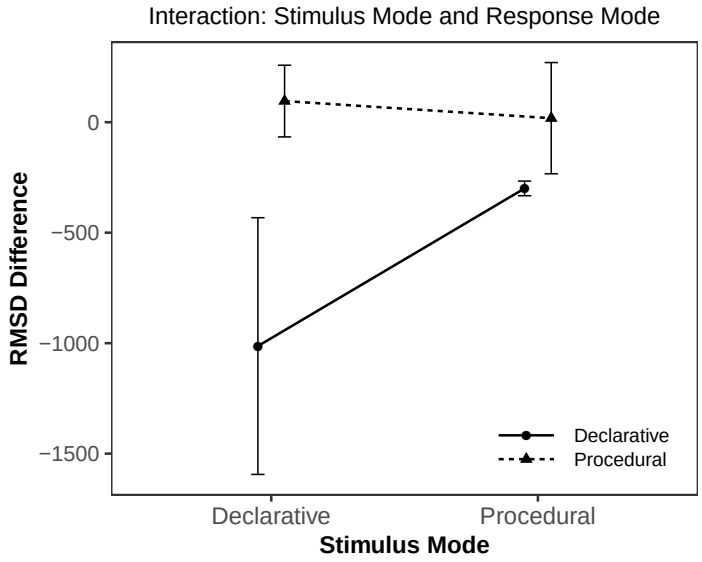

Figure 7. Differences between the median RMSD for the training region and the median RMSD for the extrapolation region, for each cell of the 2 x 2 factorial design on the first test session. Lower values indicate larger differences between performance on trained cue values and performance on extrapolation items. Error bars indicate interquartile ranges (IQRs).

With regard to the first test session, Figure 7 indicates that the difference in performance between trained cue-values and extrapolation items is larger in the declarative–declarative condition than in any other condition. A Mann-Whitney U-test on the difference between the RMSD for training items and the RMSD for extrapolation items, with the independent variable Stimulus Mode, indicated no significant difference between the declarative stimulus mode ($Mdn = -307$) and the procedural stimulus mode ($Mdn = -292$), $U = 473, p = .107$. By contrast, a Mann-Whitney U on the difference between RMSD for training items and RMSD for extrapolation items with the independent variable Response Mode indicated that the differences in performance were significantly larger with the declarative response mode ($Mdn = -313$) compared to the procedural response mode ($Mdn = 52.8$), $U = 252, p < .001$. A Kruskal-Wallis H-test, where the two variables Stimulus Mode and Response Mode were collapsed into the single independent variable Experimental Condition with four levels, revealed a significant performance difference between the conditions ($\chi^2(3) = 22.408$, $p < .001$, $\eta^2 = .294$). Pairwise

Bonferroni-corrected tests revealed that the difference between the RMSD for trained cue values and the RMSD for extrapolation items was significantly larger in the declarative–declarative condition ($Mdn = -1013$) than in the procedural–procedural condition ($Mdn = 18.2$), $U = 47, p < .001, r = .619$. It was also significantly larger than in the declarative-procedural condition ($Mdn = 95.7$), $U = 31, p < .001, r = .647$. See table S6 in the Appendix for a complete table of post-hoc tests.

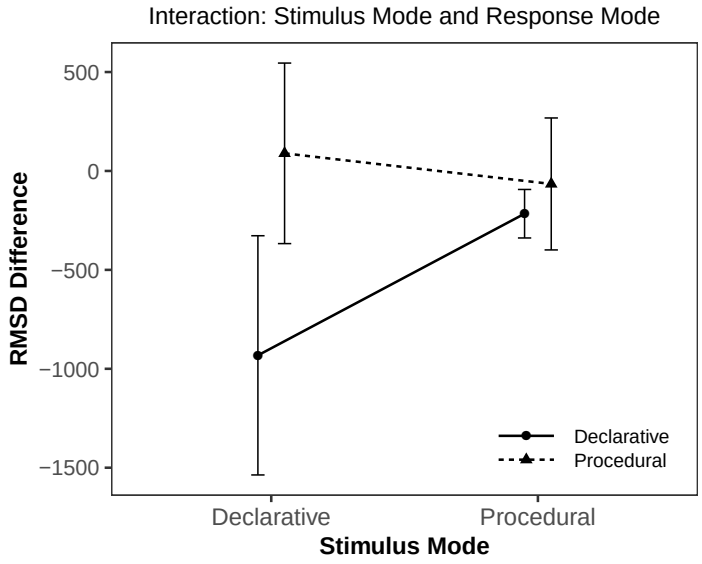

Figure 8. Differences between the median RMSD for the training region and the median RMSD for the extrapolation region, for each cell of the 2 x 2 factorial design on the second test session. Lower values indicate larger differences between performance on trained cue values and performance on extrapolation items. Error bars indicate interquartile ranges (IQRs).

With regard to the second test session, Figure 8 suggests a similar pattern as for the first test session. A Mann-Whitney U on the difference between RMSD for training items and RMSD for extrapolation items, with the independent variable Stimulus Mode, likewise indicated no significant difference between the declarative stimulus mode ($Mdn = -369$) and the procedural stimulus mode ($Mdn = -159$), $U = 472, p = .072$. By contrast, a Mann-Whitney U on the difference between training items and RMSD Response mode declarative ($Mdn = -292$) compared to procedural ($Mdn = -0.262$), $U = 362, p = .002, r = .363$. A Kruskal-Wallis H-test, where the two variables Stimulus Mode and Response Mode were collapsed into the single independent variable Experimental Condition with four levels, revealed a significant performance difference between the conditions ($\chi^2(3) = 13.896, p = .003, \eta^2 = .163$). Pairwise Bonferroni-corrected tests revealed that the difference between the RMSD for trained cue values and the RMSD for extrapolation items was significantly larger in the declarative–declarative condition than in the procedural-procedural condition $U = 67, p = .001, r = .520$. It was also significantly larger than in the declarative–procedural condition, $U = 54, p = .013, r = .526$. See table S7 in the Appendix for a complete of post-hoc tests.

As indicated in Figure 9, median differences between trained cue values and extrapolation

items were comparable for the two test sessions, with a slight decrease in difference observable in the declarative–declarative condition. A Wilcoxon Signed Rank test on the median difference between the median RMSD for trained cue-values and the median RMSD for extrapolation items was conducted separately for each cell of the 2 x 2 factorial design. The tests yielded only one significant result, indicating that that the difference between the RMSD for training items and the RMSD for extrapolation items in the procedural–declarative condition was significantly larger before the forgetting interval ($Mdn = 300$) than after the forgetting interval ($Md = 216$), $W = 36$, $p = .016$, $r = .545$. See table S8 for a complete table of Wilcoxon signed-rank tests.

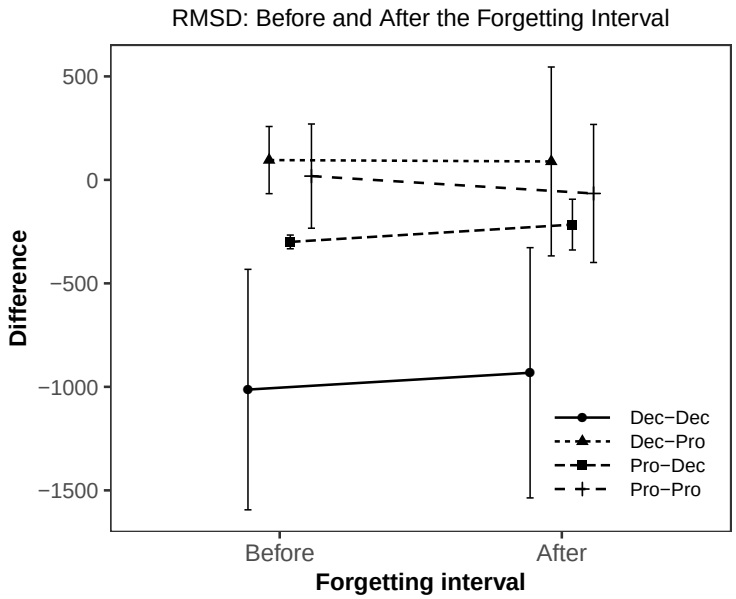

Figure 9. Median differences between median Root Mean Square Deviations (RMSDs) for trained cue-values and extrapolation items before and after the forgetting interval, for each cell of the 2x2 factorial design. The first term in the legend headings identifies the stimulus mode, and the second term identifies the response mode. Error bars represent Interquartile Ranges (IQRs).

A Kruskal-Wallis H-test, where the two variables Stimulus Mode and Response Mode were collapsed into the single independent variable Experimental Condition with four levels, revealed no significant differences between the conditions with respect to differences in RMSD between trained cue values and extrapolations before and after the forgetting interval ($\chi^2(3) = 1.898, p = .594, \eta^2 = .242$).

Modeling

In order to examine the cognitive processes involved in the function-learning task, the fits of three a priori models were evaluated.[1] The first model (”non-linear”), with two free parameters (slope and intercept), assumed that participants would extract the true underlying function form. The

[1]For all intents and purposes, the modeling procedure in the present study corresponds to the modeling procedure in Juslin et al. (2019).

second model ("linear"), with two free parameters (slope and intercept), assumed that participants would solve the function-learning task by relying on a piecewise linear approximation of the function, resulting in linear distortions similar to those illustrated in Figure 1A. The third model ("EBM"), with one free parameter, assumed that participants would rely on exemplar memory when performing the task, resulting in judgments similar to those in Figure 1B. More specifically, the model predicts that participants will assess the similarity of each probe (i.e., each cue value) to previously encountered probes with known criterion values, which are stored in memory in the form of exemplars. The estimated criterion value for each probe is determined by equation (1):

$$\hat{y}_p = \frac{\sum S(p,i) \times x_i}{\sum S(p,i)} \tag{1}$$

where $\hat{y}_p$ is the estimated criterion value for the probe p; S is the similarity of the probe to the stored exemplars; x_i is the criterion value of the exemplar i; and I is the number of stored exemplars in memory. The similarity S between a stored exemplar and the probe is given by (2):

$$S(p,i) = exp(|(p-i) \times h| \times -1) \tag{2}$$

where h is a free parameter in the range $[0.01, 10]$.

The models are a priori in the sense that, rather than fitting free parameters of the models, it is assumed that by the end of training each model has adapted to provide good approximations of the underlying function form. In other words, the free parameters—the h parameter of the exemplar model and the slope and intercept of the piecewise linear model—are fitted to capture the underlying function, thus leading to predictions like those exemplified in Figure 1B.

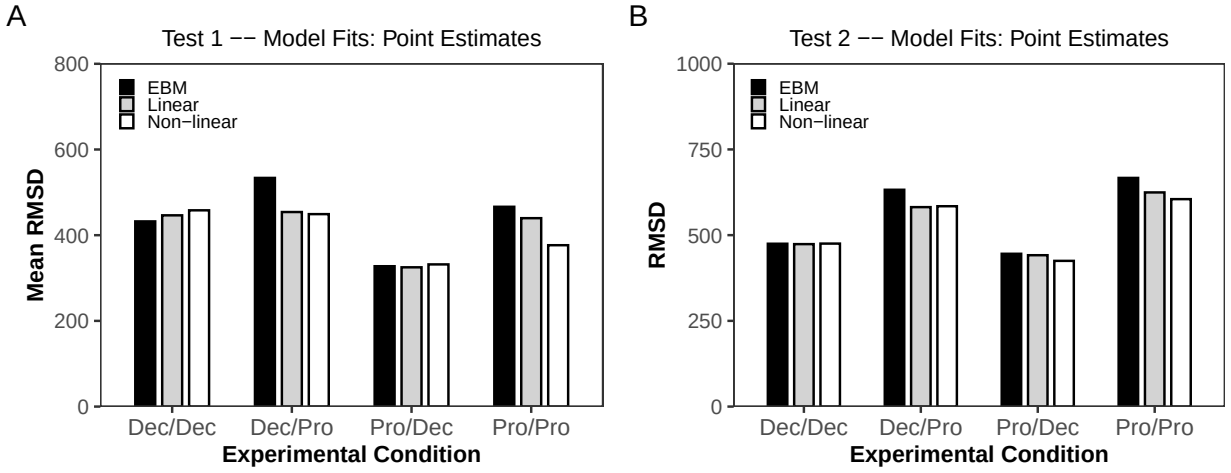

Figure 10. Mean fit of the models in terms of Root Mean Square Deviation (RMSD) between the model predictions and the judgments for each cell of the 2 x 2 factorial design. Panel A displays model fits for the first test session, and Panel B displays model fits for the second test session.

The model fits were assessed by computing the Root Mean Square Deviation (RMSD) between the predictions of the three models and the judgments by each participant. As indicated

in Figure 10, model fits were similar across all cells of the 2x2 factorial design in both the first and second test session. Contrary to predictions, the EBM model does not provide a much stronger model fit than the other models in the declarative-declarative condition.

The results of a 2x2x2x3 mixed ANOVA on the Model Fit, with the independent variables Stimulus Mode (between-subjects), Response Mode (between-subjects), Test Session (within-subjects) and Model (within-subjects) revealed significant main effects of Response Mode ($F(1,65) = 7.28, p = .009$, partial $\eta^2 = .101$) and Test Session ($F(1,65) = 12.15, p = .001$, partial $\eta^2 = .158$, as well as a significant interaction effect Response Mode $\times$ Test Session $\times$ Model ($F(2,130) = 3.22, p = .043$, partial $\eta^2 = .047$). The models fitted data from the procedural response mode better, and fitted the data from the second test session better than the first. See table S9 in the Appendix for a full report of the ANOVA.

Discussion

Historically, models of single-cue function learning have emphasized either rule-based processes (Brehmer, 1974; Koh & Meyer, 1991), exemplar-based processes (Busemeyer et al., 1997; Delosh, Busemeyer, & McDaniel, 1997), or some hybrid process involving both rule-based or exemplar-based representations (Kalish et al., 2004; McDaniel & Busemeyer, 2005). However, the results from the present study are difficult to reconcile with such models, and seem instead to point towards a conception of single-cue function learning as involving multiple, qualitatively distinct, cognitive processes, similar to current theories in the related areas of multiple-cue judgment and category learning (e.g., Juslin et al., 2008; Kruschke, 2005).

A main result of the present study is that the judgment patterns predicted by the division-of-labor hypothesis, first described and demonstrated in Juslin et al. (2019), were replicated (albeit with a few caveats). Most notably, the participants on the first test session who received both a declarative stimulus mode and a declarative response mode demonstrated a high degree of accuracy for cue values within the training region, but were unable to extrapolate in the direction of the function, displaying instead extrapolation judgments negatively correlated with the true function form. This judgment pattern corresponds well to the pattern predicted in Figure 1B. Additionally, the performance differences between trained cue values and extrapolation items were substantially less pronounced in the other cells of the 2 x 2 factorial design. By contrast, the participants who received procedural stimulus and response modes displayed an ability to extrapolate, as did participants in the two mixed conditions (i.e., participants who received either a declarative stimulus mode and a procedural response mode, or the inverse). A plausible interpretation of these findings is that the combination of a declarative stimulus mode and a declarative response mode is particularly likely to induce exemplar-based processes, which, in turn, produce judgments that capture the underlying function with great fidelity in the training region, but are marred by poor performance when extrapolating. By extension, participants who did not receive both a declarative stimulus mode and a declarative respons appeared to solve the task using a different (or several different) cognitive processes. This supports the

division-of-labor hypothesis.

However, the judgment patterns observed in the first test session for participants who received both a procedural stimulus mode and a procedural response mode did not directly replicate the corresponding results from Juslin et al. (2019). In Juslin et al. (2019), which contained a highly similar experimental condition, the data revealed pronounced linear distortions similar to those in Figure 1A; however, the judgment patterns displayed by the participants in the present study seemed to suggest that they had fully extracted the underlying function, with highly accurate performance and no hints of linear distortions. Comparing the experimental procedure in Juslin et al. (2019) to that of the present study, the salient differences relate to the design of the training and testing phases, where the present study utilized a single training phase in which each item was displayed on six different training trials in randomized order, whereas Juslin et al. (2019) used a sequence of four blocks, with each block consisting of both a training phase (containing 20 trials with outcome feedback) and a test phase (20 trials with no outcome feedback). Accordingly, it might be the case that the different outcomes in the present study compared to Juslin et al. (2019) is explained by these differences in procedure. One possibility is that the method of using a single training phase with multiple trials of each item, rather than a series of discrete blocks where training is interrupted by test phases, invites relatively stable memory traces in visual memory for the cue-criterion pairings, thus enabling the use of a process akin to exemplar memory even when the stimulus and response modes are both procedural. However, this does not explain the near-flawless extrapolation performance. A second possibility is that this task format, possibly in conjunction with the training and testing procedure, primes a cognitive process that is distinct from both exemplar memory and rule-based representations. However, the predicted linear distortions were indeed present at the second test session (Figure 4D), suggesting that the participants did rely on rule-based processes (i.e., piecewise linear approximations) after the forgetting interval. One hypothesis is thus that the first test session engaged a comparatively fragile procedural learning process that did not retain the information long enough to replicate the same performance after the forgetting interval, at which point the participants instead relied on gist memory, thus resulting in a reliance on rule-based processes. An intriguing prospect for future studies is to attempt to replicate these results, and to further investigate whether the mastering of procedural stimulus- and response modes in single-cue function learning can occur via two distinct cognitive processes.

With regard to the two mixed conditions, performance was significantly better for participants who received a procedural stimulus mode and a declarative response mode than for the participants who received a declarative stimulus mode and a procedural response mode. This might suggest that these two combinations of stimulus and response modes genuinely affect learning performance in different ways, with one displaying superior performance over the other. However, a possible limitation of the present study is that it utilized different procedural formats for the stimulus mode (a horizontal slider) and for the response mode (a bar consisting of all shades between white and black). It might be the case that one of these formats is simply more difficult than the other, despite them both being procedural. Hence, additional experiments

should examine whether similar performance differences are obtained when the same procedural format is used for both the stimulus mode and the response mode.

Another main result for the study relates to the differences in performance between the two test sessions. Notably, one explicit prediction for the second test session was that the extrapolation performance for the participants who received both a declarative stimulus mode and a declarative response mode would improve, and that their judgments would reveal linear distortions not present at the first test session (thus signifying a shift from exemplar-based processes to rule-based processes as verbatim memory traces degrade). However, the predicted linear distortions were not observed for these participants, occurring instead—as discussed—only for participants who received procedural stimulus- and response modes (Figure 4D). One possible explanation for this result is that the verbatim memory traces were not sufficiently degraded even after the forgetting interval, and that the unexpectedly strong memory retention allowed the participants who received declarative stimulus- and response modes to solve the task using exemplar-based processes. Hence, the results suggest that degradation of memory traces can shift judgments towards rule-based processes that engage piecewise linear approximations (as seen for the participants who received only procedural stimulus and response modes), but additional research is needed to delineate how, and when, degradation of memory traces impact judgments for learned functions.

Directions for Future Research

Future studies on single-cue function learning should further investigate the moderating factors for linear distortions, and the role of memory processes in distorting judgments over time. Additionally, the results from the present study hints at the presence of an additional cognitive process capable of high accuracy for both training and extrapolation regions for procedural task formats. Future studies should investigate whether these results can be replicated, and test the hypothesis that procedural task formats in single-cue function learning can engage two distinct cognitive processes.

Conclusions

In this master's thesis, data is presented that are not predicted by the currently established models of single-cue function learning. One plausible interpretation of the data is that participants utilized qualitatively distinct cognitive processes—primarily, exemplar-memory and rule-based processes, but perhaps additional processes as well—when solving single-cue function learning tasks, as suggested by the division-of-labor hypothesis.

References

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (p.157–188). Elsevier Academic Press.

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50.

Brainerd, C., & Reyna, V. (1993). Domains of fuzzy-trace theory. In *Emerging themes in cognitive development* (pp. 50–93). Springer.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*(6), 259–260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Brehmer, B. (1976). Subjects' ability to find the parameters of functional rules in probabilistic inference tasks. *Organizational Behavior and Human Performance*, *17*(2), 388–397.

Brehmer, B., Alm, H., & Warg, L.-E. (1985). Learning and hypothesis testing in probabilistic inference tasks. *Scandinavian Journal of Psychology*, *26*(1), 305–313.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Byun, E. (1996). *Interaction between prior knowledge and type of nonlinear relationship on function learning* (PhD thesis). Purdue University.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua* (PhD thesis). Princeton University.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Delosh, E. L., Busemeyer, J., & McDaniel, M. (1997). Extrapolation: The sine qua non of abstraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

Enkvist, T., Newell, B., Juslin, P., & Olsson, H. (2006). On the role of causal intervention in multiple-cue judgment: Positive and negative effects on learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 163–179.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2012). Concepts and categorization. In A. Healy & R. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (Vol. 4, pp. 599–621). Hoboken, NJ, US: John Wiley & Sons Inc.

Helversen, B. von, & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*(1), 73–96.

Hoffman, J. A., Helversen, B. von, & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of*

Experimental Psychology: General, *143*(6), 2242–2261.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT press.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298.

Juslin, P., Millroth, P., Sundh, J., & Nilsson, H. (2019). *A division of labor in single-cue function learning*. Unpublished manuscript.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. Holyoak & R. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 267–293). New York: Cambridge University Press.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072–1099.

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review*, *14*(6), 1140–1146.

Kelley, H., & Busemeyer, J. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology*, *52*(4), 218–240.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811–836.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (2005). Category learning. In K. lamberts & R. L. Goldstone (Eds.), *The handbook of cognition* (pp. 183–201). London: Sage.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1666–1684.

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General*, *131*(2), 163–193.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory and Cognition*, *28*(2), 295–305.

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory and Cognition*, *43*(2), 283–297.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 173–195.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, *35*(1), 113–138.

Olsson, A.-C., Juslin, P., & Olsson, H. (2006). Individuals and dyads in a multiple-cue judgment task: Cognitive processes and performance. *Journal of Experimental Social Psychology*, *42*(1), 40–56.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240.

Penn, D. C., & Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology*, *58*, 97–118.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*(4), 577–596.

Pinker, S. (1995). Language acquisition. *Language: An Invitation to Cognitive Science*, *1*, 135–182.

Platzer, C., & BrÃ¶der, A. (2013). When the rule is ruled out: Exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*(5), 429–441.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1–75.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63–82.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception & Psychophysics*, *18*(6), 416–422.

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1045–1064.

Appendix

Table S1. *Pairwise Bonferroni-corrected Mann-Whitney U-tests between all experimental conditions for the first test session. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	204	1	.165
Declarative–Declarative vs. Declarative-Procedural	176	1	.199
Declarative–Declarative vs. Procedural-Declarative	243	1	.165
Declarative–Procedural vs. Procedural-Procedural	173	1	.239
Declarative–Procedural vs. Procedural-Declarative	28	$< .001$	.681
Procedural-Declarative vs. Procedural-Procedural	105	.685	.33

Table S2. *Pairwise Bonferroni-corrected Mann-Whitney U-tests between all experimental conditions for the first test session, with extrapolation items excluded. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	123	.893	.240
Declarative–Declarative vs. Declarative-Procedural	91	.461	.306
Declarative–Declarative vs. Procedural-Declarative	215	1	.163
Declarative–Procedural vs. Procedural-Procedural	173	1	.239
Declarative–Procedural vs. Procedural-Declarative	270	$< .001$	.758
Procedural-Declarative vs. Procedural-Procedural	36	$< .001$	.674

Table S3. *Pairwise Bonferroni-corrected Mann-Whitney U-tests between all experimental conditions for the second test session. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	173	1	.001
Declarative–Declarative vs. Declarative-Procedural	157	1	.086
Declarative–Declarative vs. Procedural-Declarative	222	1	.197
Declarative–Procedural vs. Procedural-Procedural	126	1	.057
Declarative–Procedural vs. Procedural-Declarative	243	$< .001$	.598
Procedural-Declarative vs. Procedural-Procedural	64	.685	.535

Table S4. *Pairwise Bonferroni-corrected Mann-Whitney U-tests between all experimental conditions for the second test session, with extrapolation items excluded. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	113	.484	.290
Declarative–Declarative vs. Declarative-Procedural	110	1	.193
Declarative–Declarative vs. Procedural-Declarative	227	1	.220
Declarative–Procedural vs. Procedural-Procedural	126	1	.057
Declarative–Procedural vs. Procedural-Declarative	247	$< .001$	.622
Procedural-Declarative vs. Procedural-Procedural	49	$< .001$	.609

Table S5. *Wilcoxon signed-rank tests examining the difference between the median RMSD at the first test session and the median RMSD at the second test session, for each cell of the 2 x 2 factorial design. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*W*	*p*	*r*
Declarative–Declarative	44	.130	.249
Declarative–Procedural	17	.013	.557
Procedural–Declarative	22	.002	.674
Procedural–Procedural	18	.002	.693

Table S6. *Pairwise Bonferroni-corrected Mann-Whitney U-tests on the difference between RMSDs on trained cue values and RMSDs on extrapolation items, for all experimental conditions on the first test session. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	47	.001	.619
Declarative–Declarative vs. Declarative-Procedural	31	.001	.647
Declarative–Declarative vs. Procedural-Declarative	93	.06	.414
Declarative–Procedural vs. Procedural-Procedural	129	1	.020
Declarative–Procedural vs. Procedural-Declarative	204	.05	.450
Procedural-Declarative vs. Procedural-Procedural	112	.45	.295

Table S7. *Pairwise Bonferroni-corrected Mann-Whitney U-tests on the difference between RMSDs on trained cue values and RMSDs on extrapolation items, for all experimental conditions on the second test session. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*U*	*p*	*r*
Declarative–Declarative vs. Procedural-Procedural	67	.001	.520
Declarative–Declarative vs. Declarative-Procedural	54	.013	.526
Declarative–Declarative vs. Procedural-Declarative	105	.171	.358
Declarative–Procedural vs. Procedural-Procedural	129	1	.038
Declarative–Procedural vs. Procedural-Declarative	171	1	.170
Procedural-Declarative vs. Procedural-Procedural	127	1	.220

Table S8. *Wilcoxon signed-rank tests examining the ON-difference between the median RMSD at the first test session and the median RMSD at the second test session, for each cell of the 2 x 2 factorial design. The first term in each condition description designates the stimulus mode, and the second term designates the response mode.*

	Statistic		
Experimental condition	*W*	*p*	*r*
Declarative–Declarative	47	.170	.337
Declarative–Procedural	51	.952	.025
Procedural–Declarative	36	.016	.545
Procedural–Procedural	65	.393	.210

Table S9. *Results of a Four-way Mixed ANOVA on the Model Fit (RMSD Between Model Predictions and the Judgments) with Independent Variables Stimulus Mode (Between-Subjects), Response Mode (Between-Subjects), Session (Within-Subjects) and Model (Within-Subjects; Linear Approximation Model, Nonlinear Model, Exemplar-Based Model).*

	SS	*df*	*MS*	*F*	*p*	Partial η^2
Stimulus mode (1)	145364	1	145364	0.52	.474	.008
Response mode (2)	2036312	1	2036312	7.28	.009	.101
(1) x (2)	221103	1	221103	0.79	.377	.12
Error	18180746	65	279704			
Session (3)	1922537	1	1922537	12.15	.001	.158
(1) x (3)	203295	1	203295	1.29	.261	.019
(2) x (3)	542708	1	542708	3.43	.069	.05
(1) x (2) x (3)	18737	1	18737	0.12	.732	.002
Error	10281882	65	158183			
Model (4)	60609	2	30305	2.08	.129	.031
(1) x (4)	6476	2	3238	0.22	.801	.003
(2) x (4)	79774	2	39887	2.74	.068	.040
(1) x (2) x (4)	8711	2	4355	0.30	.742	.005
Error	1890552	130	14543			
(4) x (3)	3044	2	1522	0.41	.663	.006
(1) x (4) x (3)	5230	2	2615	0.71	.494	.011
(2) x (4) x (3)	23756	2	11878	3.22	.043	.047
(1) x (2) x (4) x (3)	6308	2	3154	0.86	.427	.013
Error	479147	130	3686			

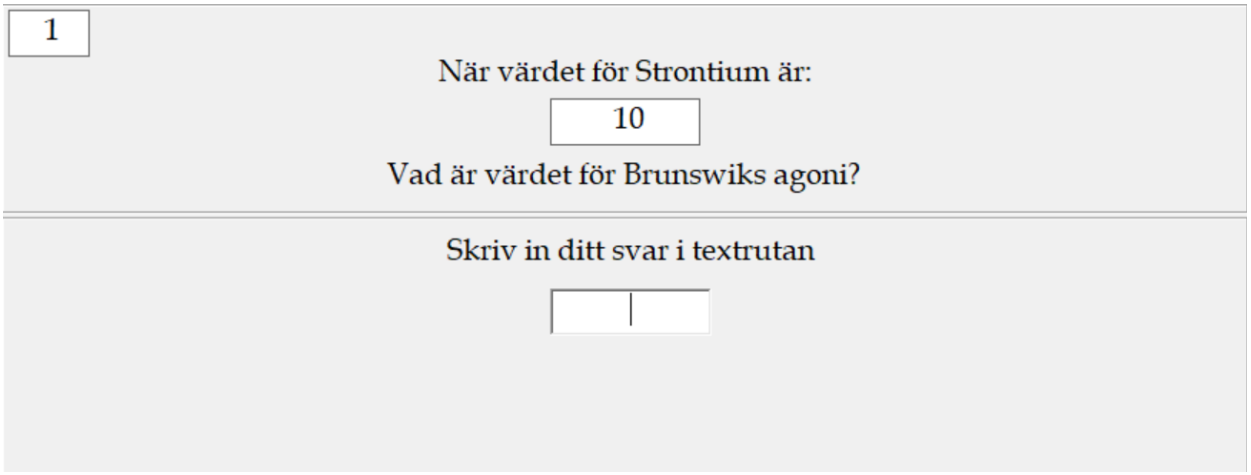

Figure S1. Screenshot from the desktop application used for the experiment. The screenshot depicts the experiment when run using a declarative stimulus mode and a declarative response mode, with Swedish language settings.

Figure S2. Screenshot from the desktop application used for the experiment. The screenshot depicts the experiment when run using a procedural stimulus mode and a procedural response mode, with Swedish language settings.