Cognition 240 (2023) 105584

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Is numerical information always beneficial? Verbal and numerical cue-integration in additive and non-additive tasks

August Collsiöö [*], Peter Juslin, Anders Winman

Department of Psychology, Uppsala University, Sweden

ARTICLE INFO

Keywords:
Judgment
Multiple-cue learning
Division of labor
Verbal
Numerical

ABSTRACT

When people use rule-based integration of abstracted cues to make multiple-cue judgments they tend to default to linear additive integration of the cues, which may interfere with efficient learning in non-additive tasks. We hypothesize that this effect becomes especially pronounced when cues are presented numerically rather than verbally, because numbers elicit expectations about a task with a simple numerical solution that can be appropriately addressed by linear and additive integration. This predicts that, relative to a verbal format, a numerical format should be advantageous for learning in additive tasks, but detrimental for learning in non-additive tasks. In two experiments, we find support for the hypothesis that a verbal format can improve learning in non-additive tasks. The division-of-labor between cognitive processes observed in previous research (Juslin et al., 2008), with cue abstraction in additive tasks and exemplar memory in non-additive tasks, was only present in conditions with numeric information and may therefore in part be driven by the use of numeric formats. This illustrates how surface characteristic of stimuli can elicit different priors about the nature of the variables and the generative model that produced the cues and the criterion. We fitted cue-abstraction and exemplar algorithms by PNP-modeling (Sundh et al., 2021). At the end of training both cue abstraction and exemplar memory processes primarily involved exact analytic processes marred by occasional error, rather than the noisy and approximate intuitive processes typically assumed in previous studies – specifically, cue abstraction was primarily implemented by number crunching and exemplar memory by rote memorization.

1. Introduction

When learning the relations between variables in a new environment, such as the functional relations between cues and a criterion and how the effect of multiple cues combine, people tend to assume linear relationships and additive integration of cues (e.g., Brehmer, 1974; Cooksey, 1996; Juslin, Nilsson, & Winman, 2009; Juslin, Nilsson, Winman, & Lindskog, 2011). That is, if one beer is good, then ten beers should, at least, be much better (monotonicity), or perhaps even ten times better (proportional linearity), and if duck liver is tasty and ice cream is good, then duck liver with ice cream should be delicious (additivity). Interestingly, most research investigating performance and cognitive processes in multiple-cue judgment involve tasks with numerical cues and a numerical criterion (Brehmer, 1994; Cooksey, 1996; Karelaia & Hogarth, 2008), although some studies have used analogue stimuli (e.g., Albrecht, Hoffmann, Pleskac, Rieskamp, & von Helversen, 2020; Hoffmann, von Helversen, & Rieskamp, 2019), or even non-metric cues (see e.g., Björkman, 1973; Castellan & Edgell, 1973; Edgell, 1983).

In everyday decision-making however, we are faced with vast amounts of non-numeric information on which we base our judgments and decisions. To exemplify, if you have a cough and are pondering if you should go to work or not, you (probably) would not measure the intensity and latency of your coughs, but rather decide based on a verbal magnitude estimation of the cough as “mild” or “severe”. At other times, quantification is not even possible, for example when making a judgment about how your first encounter with your partner’s parents will be. The cue-information you use could be information about how open minded or strict they are, probably described in terms of “very” or “not at all”.

The purpose of this article is to investigate the potential format dependence of results in the multiple-cue judgment literature by examining how the cognitive processes and learning are affected by numeric vs. verbal magnitude formats of the cues and the criterion. Do different formats for presentation of the cues and the criterion elicit different expectations (or priors) in the participants about the nature of the variables and the underlying generative process in ways that determine the nature of the cognitive processes engaged?

[*] Corresponding author at: Department of Psychology, Uppsala University, P.O. Box 1225, SE-751 42 Uppsala, Sweden.
E-mail address: august.collsioo@psyk.uu.se (A. Collsiöö).

https://doi.org/10.1016/j.cognition.2023.105584
Received 7 December 2022; Received in revised form 22 June 2023; Accepted 31 July 2023
Available online 30 August 2023

1.1. Linearity and additivity in multiple-cue judgment

We will refer to a multiple-cue judgment task as linear in Cue X_1, if – when keeping other cues constant – a certain change in X_1 always produces the same change in the Criterion Y. We refer to a multiple-cue judgment task as additive with respect to cues X_1 and X_2, if the effect of a certain change in X_1 on the Criterion Y is always the same regardless of the value of X_2 (and vice versa). A task where all the cue-criterion relationships satisfy these conditions accordingly correspond to the standard-case of the "linear additive model".

Cognitive research indicates that by default people tend to expect linear relations between cues and criterion (e.g., Brehmer, 1974; Juslin et al., 2011; Kalish, Lewandowsky, & Kruschke, 2004; Karelaia & Hogarth, 2008). Linear relationships are more available in memory and more easily learned than more complex relationships such as U-shaped relations, which may only be learned with clear and unambiguous feedback (Brehmer, 1980; Brehmer, Kuylenstierna, & Liljergren, 1974). People, likewise, typically default to considering the additive effect of each cue in isolation, independently of the values of the other available cues (Brehmer, 1994; Juslin, Karlsson, & Olsson, 2008; Karlsson, Juslin, & Olsson, 2007; Olsson, Enkvist, & Juslin, 2006), although they do also have the ability to learn non-additive relations (e.g., Juslin et al., 2008; Mellers, 1980) and configural cue patterns (Castellan & Edgell, 1973; Edgell & Castellan, 1973). People accordingly seem to approach tasks with, at least implicit, assumptions of linearity and additivity. Only if the feedback unambiguously falsifies these default assumptions, are other relations considered.

Linear, additive models are also known to be particularly robust and applicable, allowing good performance also in new and unknown environments (Brehmer, 1974, 1994; Dawes & Corrigan, 1974; Hammond, 1996; Karelaia & Hogarth, 2008). Therefore, from an adaptive and evolutionary perspective, linear additive cue integration is a plausible candidate for a default process, likely to be reinforced by everyday experience (Brehmer, 1974).

Van Dooren, De Bock, Janssens, and Verschaffel (2008) further describe the emphasis on the linear model through mathematical education and elaborate on the characteristics of mathematical education, which is suggested to produce an over-application of linearity, a "*linear imperative*", which is activated in mathematical tasks, and may impair estimation in non-linear tasks. Ebersbach, Van Dooren, Van den Noortgate, and Resing (2008), for example, show that children's understanding of exponential growth is hindered when they learn to count (a linear process), and over-apply this strategy. In addition, when presented with mathematical problems individuals tend to neglect relevant real-life knowledge suggesting non-linearity and instead adopt linear integration strategies (see e.g., Verschaffel, De Corte, & Lasure, 1994) and individuals tend to integrate based on a linear model in mathematical settings due to reliance on mathematical habits and expectations invited by the task-content (see e.g., De Bock, Van Dooren, Janssens, & Verschaffel, 2002; De Bock, Van Dooren, Janssens, & Verschaffel, 2007; Dewolf, Van Dooren, & Verschaffel, 2011). Relatedly, linear properties are overused also in probabilistic reasoning in school mathematics (see Van Dooren, De Bock, Depaepe, Janssens, & Verschaffel, 2003). In sum: the default to linear additive cue integration may be supported by cognitive constraints, its pervasive robustness for predictions in real-life tasks, as well as by mathematical education and, especially when the task is in a numeric format.

1.2. Cognitive processes in multiple-cue judgment

Much research on multiple-cue judgment has modeled the cognitive process as an interplay between rule-based processes captured by cue abstraction models and the memory- and similarity-based processes captured by exemplar models (e.g., Hoffmann, von Helversen, & Rieskamp, 2016; Juslin, Olsson, & Olsson, 2003; Juslin et al., 2008; Pachur & Olsson, 2012; Trippas & Pachur, 2019). The cue abstraction model claims that people encode the cues independently, assess their individual importance for the judgment criterion, and by default integrate these beliefs by a linear and additive cue process. The exemplar model claims that people consider the criterion values of previously observed similar exemplars, which are weighted into judgment based on the (nonlinear) similarity functions. One empirical criterion for distinguishing between the cue abstraction and the exemplar models is the ability to extrapolate the performance beyond the observed training range (DeLosh, Busemeyer, & McDaniel, 1997; Juslin et al., 2003). As illustrated in Fig. 2 in the Results section of Experiment 1, when participants rely on cue abstraction, and even if they train only on exemplars with criterion values in the range 20 to 80, because they have induced the rule-based structure, they extrapolate beyond this training range (see Fig. 2B). By contrast, when participants rely only on the similarity to concrete training exemplars, they are unable to extrapolate beyond the observed training range, that is, they do not respond with values below 20 or above 80 (Fig. 2A).

A large literature demonstrates that people shift between these two processes as a function of task properties (e.g., Hoffmann et al., 2016; Hoffmann, von Helversen, & Rieskamp, 2014; Juslin et al., 2003; Juslin et al., 2008; Karlsson et al., 2007; Olsson et al., 2006; Pachur & Olsson, 2012; Platzer & Bröder, 2013; von Helversen & Rieskamp, 2009) and properties of the decision maker (e.g., Hoffmann et al., 2014; Little & McDaniel, 2015; von Helversen, Mata, & Olsson, 2010). In multiple-cue learning tasks, the initial default process often involves attempts at explicit "problem solving", trying to abstract what cues are relevant for inferring the criterion and then to combine the cues according to a linear additive rule. As a backup strategy, to the extent that these attempts at cue abstraction prove futile, people resort to retrieving similar exemplars (configurations of cue values) by direct use of memory processes (Juslin et al., 2008: Karlsson et al., 2007; Olsson et al., 2006; see Trippas & Pachur, 2019, for results qualifying this claim).

In the present study, we explore how the relative support for these two models is affected by whether the tasks are presented in a numerical or verbal format. In the General Discussion, we return to an elaborate discussion of limits on identifiability of the processes and representations in tasks like these, the scope of generalization, and how the results can be interpreted in models that assume blends of the two processes (Bröder, Gräf, & Kieslich, 2017).

The described research on multiple-cue judgment explains the inclination to assume linear additive cue-criterion relations by claiming that cue abstraction processes are strongly constrained by working memory resources to primarily induce linear cue-criterion relations and additive inter-cue relations. For example, the default to infer linear relationships may, at least in part, be driven by that people often only consider two X,Y coordinates at a time in working memory, which naturally supports detection of the sign of a linear relationship, but provides little support for identifying nonlinear relations (Juslin et al., 2008). Because of constraints on working memory capacity, people may have difficulty with interpreting and responding differently to the value of X_1, depending on the value of other cues (e.g., X_2), as required to capture non-additive inter-cue relations when sequentially attending to the cues (Juslin et al., 2008). The human preference for the "simplicity and elegance" of the linear model may thus in part also derive from very basic information processing constraints.

1.3. Analytic or intuitive cognitive processes

Based on the Precise-Not-Precise (PNP)-model (Sundh, Collsiöö, Millroth, & Juslin, 2021), we further distinguish between two different interpretations of cue abstraction and exemplar memory processes, respectively. The PNP model draws on a proposal by Brunswik (1956), that analytic processes are typically precise but occasionally marred by large errors (a leptokurtic distribution) and that intuitive processes are approximate and ubiquitously perturbed by normally distributed noise.[1] With Analysis, the response distribution is concentrated at the deterministic execution of the algorithm (see the histograms in Fig. 4 under 2.2.3 Cognitive modeling), while with Intuition the processes are always perturbed by a random noise that yields a normal distribution (the continuous function in Fig. 4). Good fit of a cue abstraction model may thus either imply explicit and analytical rule-based integration according to explicit formulae (i.e., crunching an explicit equation in working memory), or the intuitive additive weighting of separate rules for cue-criterion relations that is assumed in much of the multiple-cue judgment research in the Brunswikian tradition (Brehmer, 1994; Karelaia & Hogarth, 2008; Sundh et al., 2021).

Likewise, the good fit of an exemplar-based memory model may either signify rote memorization of exemplars, or exemplar-based inference that involves similarity-based weighting of exemplars (i.e. the standard interpretation of good fit with the Generalized Context Model by Nosofsky, 2011; see Collsiöö, Sundh, & Juslin, 2023; Izydorczyk & Bröder, 2021). A response based on rote-memory, as when you recall the year of your birth, (almost) always gives the same correct result, leading to an extremely leptokurtic distribution. Similarity-based inferences from known exemplars, as when you estimate the price level of a new restaurant based on the price-levels of similar known restaurants, is more likely to produce a variable output from time to time that is better described by a Gaussian distribution. As detailed in the Method section below, the distribution of judgment errors and PNP modeling (Sundh et al., 2021) allow us to distinguish between the analytic and intuitive applications of these two cognitive processes. Most previous research on multiple-cue judgment has made the implicit assumptions that the good fit of cue abstraction models mainly involves intuitive cue integration (rather than "explicit number crunching"), and likewise that the good fit of exemplar models mainly involves similarity-based weighting of exemplars (rather than rote memorization and the retrieval of individual exemplars). With the PNP-model, in this article we can now subject these assumptions about the processes to stringent empirical test.

1.4. What are the effects of numeric formats on the task priors?

Although there are important exceptions (e.g., pictorial cues in Albrecht et al., 2020; Hoffmann et al., 2019), it seems fair to conclude that most of the studies of multiple-cue judgment with metric cues have relied on a numerical format (Brehmer, 1994; Cooksey, 1996; Karelaia & Hogarth, 2008). The claim for the widespread use of linear additive cue integration may thus be over-stated or premature, given that many studies use a numeric format that in itself may elicit a "linear imperative" (De Bock et al., 2002; De Bock et al., 2007; Dewolf et al., 2011; Van Dooren et al., 2008; Verschaffel et al., 1994). A numeric format may affect the participants' expectations (or "priors") about the task in two different ways. First, the choice of numbers may elicit the expectation that the variables so represented have a metric and cardinal nature. Second, the numbers may elicit the expectation that the cues and criterion are governed by an underlying simple equation that can be induced (”cracked") by intense efforts at problem solving. Both of these factors are likely to encourage linear additive cue integration, because – as we have seen – mathematical education supports linear thinking as a default rule that is applied whenever it is triggered by mathematical cues in the context (De Bock et al., 2002; De Bock et al., 2007; Dewolf et al., 2011; Van Dooren et al., 2008; Verschaffel et al., 1994).

In order to investigate if a numeric format promotes increased reliance on linear additive rules in multiple-cue tasks, we need an alternative format to compare the numeric format to. In this article, we compare a numeric format with a verbal magnitude format. In addition to the research on mathematical education, there are several lines of research in Psychology that suggest differences between numeric and verbal formats. When people integrate verbal magnitude phrases, in comparison to numerical information, they rely more on interpreting the information and adapting their strategy to the context, rather than to use context-independent rules (Liu, Juanchich, Sirota, & Orbell, 2020a). This partially explains why numbers require more effort to process, as the correct interpretation of the contextual meaning of a numerical magnitude is not necessarily provided by the context (Childers & Viswanathan, 2000).

Additionally, numeric measures are suggested to invite rule-based reasoning while verbal measures invite associative and intuitive thinking (Windschitl & Wells, 1996), and verbal information is recalled better than numerical information (Scammon, 1977). Furthermore, individuals prefer to receive and communicate with verbal statements when information is unreliable or imprecise and with numerical statements when information is reliable and precise (see e.g., Budescu & Wallsten, 1987; Wallsten, Budescu, Zwick, & Kemp, 1993). This suggests that individuals infer suitable processing strategies from the format of the information.

Schkade and Kleinmuntz (1994) show how numbers, relative to verbal magnitude information, resulted in more compensatory actions (e.g., trading off attributes) relative to non-compensatory actions (e.g., elimination of attributes or relying on a cut-off value) and more arithmetic and summary actions, where the latter constitute for example aggregating attribute values. Verbal formats lead to relatively more effort spent on acquisition of information, rather than integration, and increased alternative based search, that is reading all information for one alternative at a time rather than investigating individual cues over all alternatives (Schkade & Kleinmuntz, 1994; Stone & Schkade, 1991). These results again indicate increased tendency for linear integration (which is dependent on arithmetic and summary actions as well as focus on one cue at a time) when magnitudes are numeric rather than verbal, and a greater role for memory-based processes with verbal information (which are dependent on effort spend on acquisition of information related to an alternative).

Windschitl and Wells (1996) suggest that while a numeric format should be advantageous in situation where rule-based reasoning is beneficial, verbal formats should be superior in environments where people need to apply strategies that are not rule-based (e.g., memory-based ones). Together this suggests that one could expect verbal magnitude formats to, in terms of the PNP-model, invite intuitive processes as the format invites associative processes and an expectation about inherent variability in the underlying processes.

1.5. Purpose of the present study

The predictions for the experiments were derived from two assumptions and a new hypothesis. The first assumption is that of a *division of labor* between cognitive processes (Juslin et al., 2008). People have difficulty with capturing non-additive cue criterion relations with their controlled rule-based thinking (cue abstraction). In tasks that require non-additive integration of several cue-criterion relations, they therefore have to shift to exemplar-memory processes (Hoffmann et al., 2014, 2016; Juslin et al., 2003; Juslin et al., 2008; Karlsson et al., 2007; Pachur & Olsson, 2012; Platzer & Bröder, 2013; von Helversen & Rieskamp, 2009).

The second assumption is that of a *rule bias* (Ashby & Maddox, 2005; Juslin et al., 2008): people are inclined to start the learning of a novel task in a "problem solving mode". They initially try to induce the rules that connect the individual cues to the criterion (cue abstraction), turning to less explicit and more effortful learning strategies, like exemplar memory, only if the cue abstraction fails to allow satisfactory

[1] Note that this definition of analysis and intuition is different from the dual-systems definition (e.g., Evans, 2008; Evans & Stanovich, 2013) of intuition and analysis.

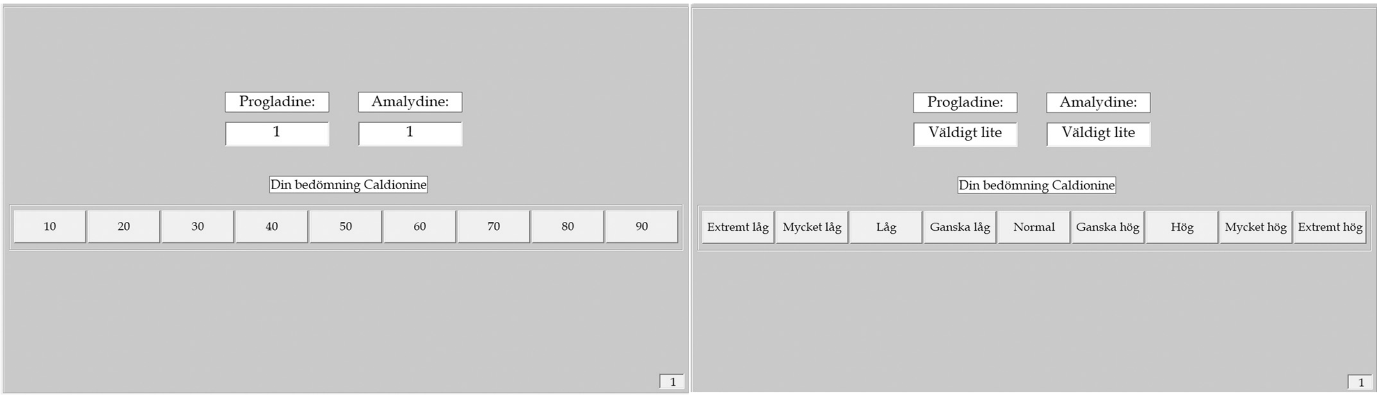


Fig. 1. Task Appearance for the Numeric Format and the Verbal Magnitude Format.
Note: The cue values are presented in the boxes with the headers "Progladine" and "Amalydine". Below this it reads "Your judgment of Caldionine" and then follows the 9-step response scale.

performance (or, alternatively, because the task becomes over-learned and potentially automatized).

The new hypothesis is that words, in comparison to numbers, should decrease the rule bias, the initial inclination to attempt at explicit cue abstraction encouraging an immediate or faster transition to exemplar memory. Numbers will invite linear additive rules improving the learning in a linear-additive task, but delaying learning in a non-additive task, where the participants are required to switch to an exemplar-memory based strategy (Juslin et al., 2003; Juslin et al., 2008). Research on mathematical learning accordingly suggest over-application of linear rules when numbers are present (De Bock et al., 2002; De Bock et al., 2007; Van Dooren et al., 2008; Verschaffel et al., 1994) and that verbal information invites relatively more associative reasoning (e.g., Liu et al., 2020a; Liu, Juanchich, Sirota, & Orbell, 2020b; Windschitl & Wells, 1996).

The prediction therefore is that, relative to numbers, a verbal magnitude format should impede learning in an additive task (where cue abstraction is viable and efficient), but speed up learning in a non-additive task (where cue abstraction is difficult and a shift to exemplar memory is required). These assumptions also predict that at the end of training (in a test phase) regardless of the cue-criterion format most participants should rely on exemplar memory in a non-additive task, while many of the participants should use cue abstraction in an additive task, especially if they address the task with numbers that invite use of cue abstraction. By applying the PNP-model to the test phase data, we can empirically test if the cognitive processes reveal the hallmarks of analytic vs. intuitive cognitive processes and explore if the effects of a numerical format are mediated by a shift to analytic cognitive processes.

2. Experiment 1: Verbal and Numerical Formats in Multiple-cue Judgment Tasks

2.1. Method

2.1.1. Participants

Eighty participants[2] (63 females, 16 males and 1 non-binary individual) ranging in age from 18 to 75 ($M = 26.33$, $SD = 8.56$) were recruited through public advertisement at various places at Uppsala University. Compensation was awarded in the form of a cinema voucher or (for students at the Department of Psychology) course credit.

2.1.2. Design

The experiment was a 2 × 2 between-subjects factorial design with format (verbal or numerical magnitude) and task (additive or non-additive cue-criterion relation) as the independent between-subject variables. The data from the training phase of the experiment was analyzed as a 2(format) x 2(task) x 10(training blocks) mixed factorial design with repeated measurement across training blocks. The dependent measure was the participants' judgments of the criterion and the accuracy of the judgments measured by the root mean square error (RMSE) between the judgment and the criterion. The responses and criteria in the verbal magnitude format condition were converted to their numerical counterparts in order to facilitate parametric testing and comparison across conditions.

2.1.3. Material

The participant's task was to judge an individual's blood concentration of the fictitious hormone Caldionine based on information about the amount of the two other fictitious hormones, Progladine and Amalydine, in the individual's urine. Both Progladine and Amalydine could take five values (1, 2, 3, 4, 5 or very little, a little, average, a lot, very much depending on the format condition).[3] A 5 × 5 factorial combination produced 25 items.

The criterion, Caldionine, could take nine values (10, 20, 30, 40, 50, 60, 70, 80, 90 or extremely low, very low, low, somewhat low, normal, somewhat high, high, very high, extremely high depending on condition). Criterion values were created from the numerical cue values, and then mapped to their verbal counterpart for the verbal magnitude format conditions. The computer program used for stimulus presentation by default saved response times. The participants answered on a fixed scale (see Fig. 1 above for an illustration). A full list of the numerical items for the additive and non-additive conditions are presented in Appendix A.

The normative additive and non-additive functions for inferring the criterion C (called Caldionine) from the cues P (Progladine) and A (Amalydine), respectively, were:

$$C(additive) = 50 + 10P - 10A \tag{1}$$

$$C(non - additive) = 50 + 10(P - 3)(A - 3) \tag{2}$$

2.1.4. Procedure

The participants conducted the experiment in separate computer booths at the Department of Psychology at Uppsala University under

[2] Note that six participants were removed as outliers for the analysis of performance during the test phase and five participants were removed as outliers for the analysis of performance during the training phase due to deviating >1.5 interquartile ranges from Q_3 for RMSE (i.e. for having extremely high RMSE).

[3] The experiment was carried out in Swedish and verbal cue- and criterion values were presented in Swedish.

scrutiny of an experiment leader. They were randomly assigned to one of the between-subject cells. The experiment consisted of a training phase and a test phase. The training phase included 23 of the 25 items, items with extreme values (10 and 90) were either all excluded (additive task) or some excluded (non-additive task) in order to investigate the use of exemplar-based memory.[4] The participants conducted 10 training blocks, each with 23 items presented in an individually randomized order, resulting in 230 training trials per participant. The test phase consisted of two blocks with all 25 items, randomized for each block. Trials were presented one at a time and the participants recorded their estimate for each trial before moving on to the next. After each training trial participants received feedback with the correct concentration of Caldionine. After the test phase, participants wrote a short description of how they solved the task.

2.1.5. *Cognitive modeling: The PNP-model*

To investigate if the participants solved the task by using a deterministic and analytic process (e.g., number crunching or retrieving a known fact) or an approximate, intuitive process perturbed by noise, we used the PNP model (Sundh et al., 2021). The PNP model assumes that intuitive processes involve a homogenous Gaussian noise around the output of a cognitive algorithm, whereas analytic processes yield leptokurtic (spiked) distributions, effectively sampling from two distributions: **i**) error-free application of the algorithm and **ii**) occasional erroneous execution of the algorithm. If B is a Bernoulli random variable with probability λ that an error occurs in execution of the algorithm, each estimate y given a cognitive process (function) $g(\mathbf{x}|\boldsymbol{\theta})$ is defined by

$$y|(B=b) = \begin{cases} g(\mathbf{x}|\boldsymbol{\theta}) + N(0,\sigma^2), b = 1 \\ g(\mathbf{x}|\boldsymbol{\theta}) + N(0,\tau^2), \quad b = 0 \end{cases} \tag{3}$$

For an intuitive process we have $\lambda = 1$ and ubiquitous Gaussian noise perturb the output of the model (as assumed in most statistical modeling). For an analytic process, λ is a small number. For technical reasons, it is prudent to introduce a very narrow Gaussian tolerance around the error free responses, as describe by the fixed parameter τ (see Sundh et al., 2021 for details).

The PNP model was fitted to individual participant data from the test phase with three cognitive process functions $g(\mathbf{x}|\boldsymbol{\theta})$. To capture additive and non-additive cue abstraction (see Eq. 1 and Eq. 2 above for the normative values for constants) we fitted two rule-based models to data:

$$g(\mathbf{x}|\boldsymbol{\theta}) = \alpha + \omega_1 {}^*P - \omega_2 {}^*A, \tag{4}$$

$$g(\mathbf{x}|\boldsymbol{\theta}) = \alpha + \omega_1 {}^*(P - \omega_2)^*(A - \omega_2). \tag{5}$$

Each model gives a prediction of the participant's response for the value of the criterion (Caldionine) where α represents the intercept and P (Progladine) and A (Amalydine) represent the cues. In the additive rule-based model (Eq. 4) ω_1 is the weight given to cue P and ω_2 is the weight given to cue A. In the non-additive model (Eq. 5) ω_1 is the weight given to the product of the cues after subtraction of the ω_2 constant, $(P - \omega_2)^*(A - \omega_2)$, where ω_2 represents the constant to be subtracted from the cue-values before multiplication. α, ω_1 and ω_2 were modeled as free parameters. The memory-based processes were captured by the Generalized Context Model (see Nosofsky, 1984, 2011), applied to a continuous criterion (see e.g., Juslin et al., 2003),

$$g(\mathbf{x}|\boldsymbol{\theta}) = \frac{\sum_{j=1}^{23} exp\left(-\beta \sum_{i=1}^{2} \left|x_i - x_{ji}^*\right|\right)^* C_j}{\sum_{j=1}^{23} exp\left(-\beta \sum_{i=1}^{2} \left|x_i - x_{ji}^*\right|\right)}, \tag{6}$$

where $g(\mathbf{x}|\boldsymbol{\theta})$ represents a weighted average of the criterion C_j of each of the 23 exemplars from the feedback training based on the relative similarity of the cue-values ($x^*_{j1} \ldots x^*_{j2}$) of the exemplars to the corresponding cue-values ($x_1 \ldots x_2$) of the probe, that is the test-item. β defines how much the relative similarity between an exemplar and the test-item affects the weight put on each exemplar. When $\beta = 1$ all exemplars are thus weighted in accordance to their exponential similarity to the test-item, whereas higher values of β results in more weight on the exemplar(s) most similar to the test-item relative to less similar exemplars. Lower values of β thus result in relatively more weight on less similar exemplars and with $\beta = 0$ all exemplars are weighted equally. We fitted both a configural and a non-configural version of the exemplar-based memory model, because in the non-additive environment only the relation between the cue-values (and not their position) is relevant for the value of the criterion (i.e., the task setting is non-configural).[5]

We fitted all three models with λ as a free parameter.[6] A low λ –value indicates an analytic process, and a high λ –value indicates an intuitive process. For example, a low λ and parameters $\alpha = 50, \beta_1 = 10, \beta_2 = 10$ in the additive task (Eq. 1) indicate analytic execution (number crunching) of Eq. 4, marred by occasional errors in the computations.[7] In other words, most responses are perfect executions of Eq. 1 with the correct parameters.

2.1.5.1. *Model fit.*

Parameters were estimated by maximum likelihood estimation[8] and the Bayesian Information Criterion (BIC) was used to identify the best model fit for each individual (see Raftery, 1995). BIC contributes no information on the absolute fit, so we also report adjusted R^2 of the models with the best relative fit.[9] If we identify the correct model, all systematic variance in data should be accounted for and all residual noise should be random. We therefore also report a Saturation Index (SI) defined by

$$SI = R^2_{adj} \Big/ \rho, \tag{7}$$

where ρ is the reliability coefficient, or proportion of systematic variance in data, as estimated by the test-retest reliability when participants perform each judgment twice. SI will approach 1 if the model accounts for all systematic variance in data, while a low SI suggests that there is non-trivial residual systematicity that the model fails to explain (see Sundh et al., 2021 for further discussion on this measure). That is, if for a participant R^2_{adj} is equal to 0.6 and test-retest reliability for the participant is 0.6 the model captures all systematic variance. However, if test-

[4] See Appendix A for a full list of items, including items excluded from training.

[5] For the configural exemplar-based model similarities were calculated according to Eq. 6. For the non-configural exemplar-based model, because cue-order is immaterial to the process, the cue pairs for each item were first sorted so the lower value was viewed as cue1, regardless of original position and similarities were then calculated according to Eq. 6. That is the item with $\text{cue}_1 = 5$ and $\text{cue}_2 = 1$ was after reordering identical to the item with $\text{cue}_1 = 1$ and $\text{cue}_2 = 5$.

[6] The Raw- and processed data analyzed in the present paper, as well as MATLAB code for use of the PNP model, are available here: https://osf.io/qx6gt/

[7] Given the 9-step scale we do not expect λ –values above 0.89, given that participants should by chance answer precisely in 1 out of 9 times. As the participants responded by selecting one of nine alternative answers, the tolerance τ was set so that only the selection of the correct alternative corresponds to a perfectly correct response.

[8] The PNP model uses the function *fminsearchbnd* (D'Errico, 2022) for maximum likelihood estimation.

[9] The standard version of adjusted R^2 was used, $R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p}$, where n equals sample size (number of items) and p equals number of explanatory variables.

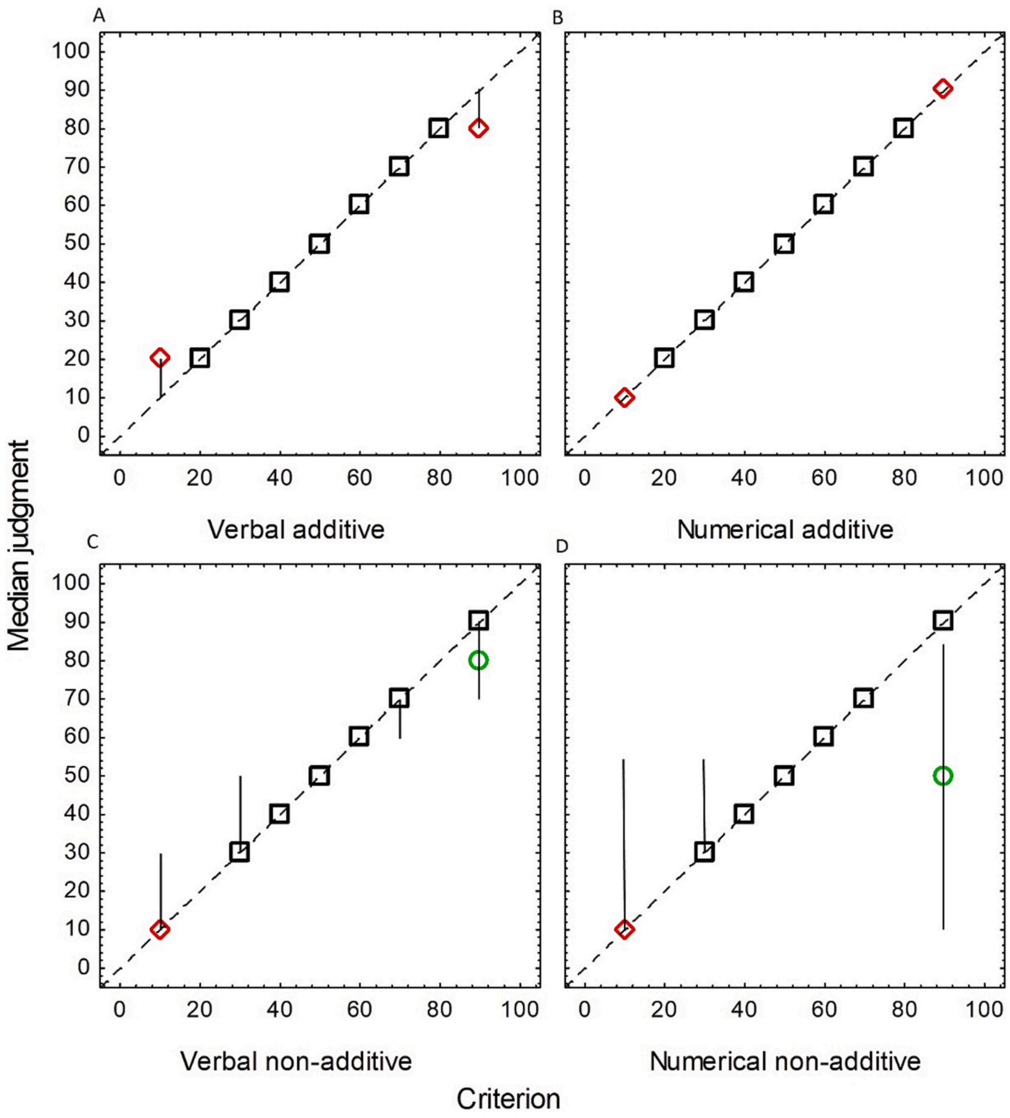


Fig. 2. Median Judgments in Experiment 1.

Note. This figure displays the median judgments for each stimulus in the test phase as function of the criterion value in each of the four cells of Experiment 1. The identity line represents perfectly accurate judgments. Diamond items (red) are extreme items requiring a response outside of the training range given a configural exemplar based memory strategy. Circle items (green) are extreme items requiring a response outside of the training range for both a configural and a non-configural exemplar based memory strategy in the non-additive tasks. Solid lines indicate interquartile ranges. Multiple points and lines overlap.

retest reliability of the participant would have been 0.9 there would be systematic variance in the data not accounted for by the model. Likewise, if R^2_{adj} is larger than the test-retest reliability the model accounts for more variance than the actual systematic variance in the data, indicating overfitting.

2.2. Results

Fig. 2 presents the median judgments for each item in the test phase in each of the experimental cells of Experiment 1. The median judgments in the verbal additive task yield the pattern implied by exemplar memory, with poorer judgments for the extrapolation items. With the numerical additive task, the median judgments reproduce the pattern with accurate extrapolation implied by cue abstraction. The lower panels for the non-additive tasks are more ambiguous. In the low range, the participants extrapolate, suggestive of cue abstraction. In the high range, however, the participants fail to produce the extreme judgments required for the item not presented during training, as predicted by exemplar memory.

One hint to a possible explanation for these puzzling results is to note that in the non-additive (but not the additive) task, the order of the cues is immaterial. In contrast to the additive task, the cues [X, Y] always imply the same criterion value as cues [Y, X], so in the non-additive task participants are well advised to ignore the order of the cues. An exemplar model with non-configural coding assumes that exemplar [X, Y] is perceived to be identical to exemplar [Y, X], and yields asymmetric predictions similar to the data illustrated in the lower panels of Fig. 2.[10] As will be clear in the section on cognitive modeling below, however, also taking this possibility into account, the results show that the models considered in this article are insufficient to fully account for the data from the numerical non-additive task.

[10] Ignoring the order of the cues, in the low range, there exists perfect twins [5,1] to the items that require extrapolation [1,5] in the training set, that has the required criterion value of 10, so retrieval of this latter item produces the correct judgment. By contrast, in the upper range there exists no such (non-configural) twin to the new item that requires an equally extreme response in the training range (criterion 90). See all numeric items in Appendix A.

Table 1
Support for each Factor of a 2x2x10 Mixed Factorial BANOVA based on the Training Phase.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Block	0.263	0.263	4.698e − 4	5.013e − 87	9.372e + 82
Task	0.263	0.263	1.466e − 4	9.496e − 15	1.544e + 10
Format	0.263	0.263	0.053	0.237	0.224
Block ✻Task	0.263	0.263	0.983	4.777e− 4	2058.392
Block✻Format	0.263	0.263	0.015	0.732	0.021
Task ✻ Format	0.263	0.263	0.693	0.054	12.786
Block ✻ Task ✻ Format	0.053	0.053	0.016	0.014	1.174

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF_{incl}) is the primary factor of concern showing the evidence for including a factor in the final model.

2.2.1. *Performance during training*

We conducted a 2x2x10 mixed factorial BANOVA with RMSE as the dependent variable and format (verbal/numerical) and task (additive/non-additive) as between-subject independent variables and training block (1−10) as the within-subject variable. The best supported model includes a main effect of block, format, and task, as well as an interaction between block and task and format and task ($BF_M = 37.968$, $BF_{10} > 10^{96}$, $BF_{2nd\ best\ model} = 2.864$).[11] The support for each factor is presented in Table 1. There is extreme evidence for main effects of block and task and for an interaction between block and task. Strong evidence is found for an interaction between format and task and inconclusive evidence for a potential three-way interaction between format, task, and block. The results are illustrated in Fig. 3, which highlight the interaction between format and task: In the additive task the verbal format (*Mean RMSE* = 8.753, *SD* = 4.200) is detrimental for performance relative to the numerical format ($M = 5.650$, $SD = 2.301$), while in the non-additive task the verbal format ($M = 14.412$, $SD = 4.908$) is advantageous in relation to the numerical format ($M = 16.885$, $SD = 3.295$).

2.2.2. *Performance at test*

A factorial BANOVA with RMSE from the test phase (including both old items from the training phase and new extrapolation items) as the dependent variable, revealed no evidence for a main effect of format ($BF_{incl} = 0.262$), extreme evidence for a main effect of task ($BF_{incl} =$ 27,894.988), and evidence for an interaction between format and task in the test phase of the experiment ($BF_{incl} = 3.995$).[12] The interaction pattern is the same pattern as in the training phase, with the participants in the numerical additive task ($M = 2.073$, $SD = 2.680$) performing better than the participants in the verbal additive task ($M = 5.239$, $SD = 2.973$), whereas the participants in the verbal non-additive task ($M = 9.782$, $SD = 6.959$) had a better performance than the participants in the numerical non-additive task ($M = 13.934$, $SD = 8.768$).

The interaction is explained by the new extrapolation items in the test phase, as verified by one-sample Bayesian *t*-tests on the difference in RMSE between new items and training items. There is no evidence for a performance difference between the old items and the extrapolation items in the numerical additive task (Fig. 2B, *Mean difference in RMSE* = −0.330, *SD* = 3.469, $n = 17$, $BF_{10} = 0.267$, $BF_{01} = 3.750$) and very weak evidence for a difference in the verbal additive task (Fig. 2A, $M = 4.419$, $SD = 8.557$ $n = 17$, $BF_{10} = 1.492$). There is, however, medium evidence for poorer performance for the new items in the verbal non-additive task (Fig. 2C, $M = 8.919$, $SD = 14.740$, $n = 20$, $BF_{10} = 3.873$) and extreme evidence in the numerical non-additive task (Fig. 2D, $M = 29.508$, $SD = 19.706$, $n = 20$, $BF_{10} = 9019.650$).

The poorer performance in the verbal-additive and both non-additive tasks indicates inability to extrapolate, suggestive of exemplar memory. This is investigated further under 2.2.3 Cognitive modeling below.

2.2.3. *Cognitive modeling*

We categorized a model as supported for a participant, if the BIC-difference between that model and all other models[13] were < -2 (see Raftery, 1995). If not, the participant was left "Uncategorized". A null model assuming that the response is the mean response for all trials[14] had poor fit for all but three participants. The SI medians in Table 2 for the best-fitting model for each participant suggest that the models account for almost all of the systematic variance in the additive tasks (SI ~ 1), whereas SI is lower in the non-additive tasks and especially with a numeric format. For the numerical non-additive task, the median best fitting model only accounts for 67.7% of the systematic variance. Evidently, for many participants in this condition, the cognitive process is not well captured by the models considered here.[15]

In all conditions, the fitted λ is far below 1 that signifies a Gaussian distribution and intuition (Sundh et al., 2021). Surprisingly, λ was low also when the exemplar model was the best fitting model, although somewhat higher with exemplar memory than with cue abstraction (median $\lambda = \mathbf{0}.18$ for exemplar models vs. median $\lambda = \mathbf{0}.02$ for cue abstraction; $BF_{10} = 267.891$, $n = 76$, Bayesian Mann-Whitney test). Fig. 4 illustrates residual distributions across all participants best fitted by a cue abstraction model (Panel A) and an exemplar model (Panel B), after subtracting the model predictions from the responses. In both cases, the residuals are distinctly leptokurtic and deviate from a normal distribution, indicative of an analytic process. In these simple tasks, at the end of training, not only cue abstraction, but also exemplar memory, takes an analytic form: rote-memorization of individual exemplars.

Median parameter estimates for the best fitting models appear in Appendix C. The median parameter estimates for the coefficients of the cue abstraction models (see Eqs. 4 & 5) coincide exactly with the correct constants in the tasks (50, 10, 10 & 50, 10, 3, respectively). The typical response thus coincides with analytic execution of these equations (or a process invariably producing the same response). Bayesian Mann-Whitney tests revealed no evidence for or against main effects on λ of format (verbal format median $\lambda = 0.16$ vs. numeric format median $\lambda = 0.06$; $BF_{10} = 0.699$, $BF_{01} = 1.430$, $n = 76$, Bayesian Mann-Whitney test) or task (additive task median $\lambda = 0.044$ vs non-additive task median $\lambda = 0.16$; $BF_{10} = 0.967$, $BF_{10} = 1.034$, $n = 76$, Bayesian Mann-Whitney test).

[11] See Appendix B for a full presentation of model comparisons.

[12] See Appendix B for a full presentation of model comparisons.

[13] Note, model comparison was focused on if participants were best fit by the additive cue abstraction model, the non-additive cue abstraction model, any exemplar-based model or the null model. Hence BIC of the best fitting exemplar-based model (regardless if it was the configuratory or non-configuratory version) was compared to the BIC of the other models for calculation of BIC-difference and best-fitting model for each participant.

[14] If the model with best fit predicted the same response for all trials, the participant was also categorized as being best described by the null-model.

[15] There is bimodality in the distributions for the Adj R^2 and the SI for the numeric non-addditive task. The non-additive cue abstraction model provides quite good fit for some participants. The poor fit is observed for some of the participants that were best fitted by the exemplar model (especially the standard configuratory one).

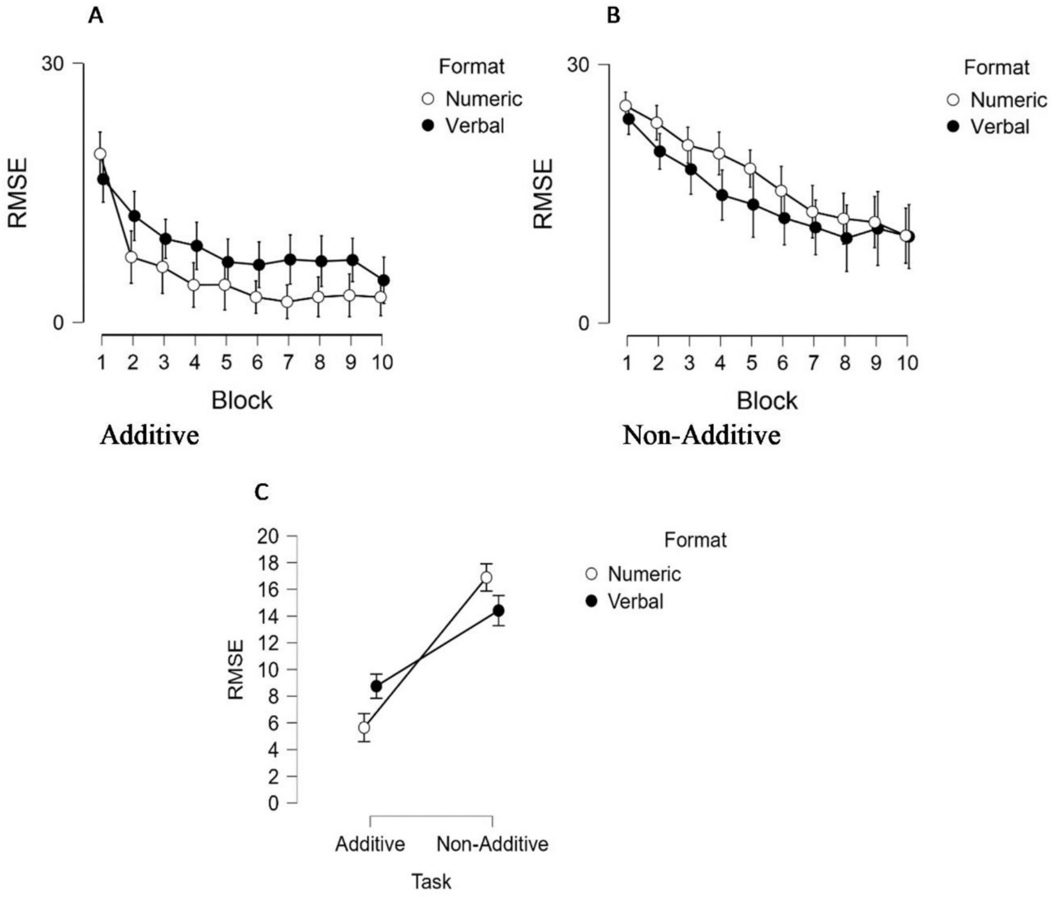


Fig. 3. Effects of Cue-Criterion Format and Cue-Criterion Relationship on Performance (RMSE) during Training.
Note. Effects of cue-criterion format and cue-criterion relationship on performance (RMSE) during training for each block of training in the additive task (Panel A) and the non-additive task (Panel B). The effects of cue-criterion format and cue-criterion relationship on performance (RMSE) during training (Panel C). Error bars are 95% credible intervals.

Table 2
Median Adjusted R^2 and SI (Saturation Index) with Interquartile Index for the Best Fitting Model (Determined by BIC) for Each Individual Grouped by the Conditions in Experiment 1.

			Task		
		Index	Additive	Non-additive	Main effect (Format)
Format	Numeric	R^2	0.995 [0.916: 1.000]	0.585 [0.117: 0.945]	0.923 [0.392: 1.000]
		SI	0.999 [0.980: 1.000]	0.677 [0.336: 0.990]	0.990 [0.600: 1.000]
	Verbal	R^2	0.931 [0.816: 0.978]	0.854 [0.386: 0.932]	0.894 [0.605: 0.970]
		SI	0.990 [0.951: 0.998]	0.938 [0.740: 0.999]	0.979 [0.867: 0.998]
Main effect (Task)		R^2	0.964 [0.854: 1.000]	0.678 [0.328: 0.963]	0.911 [0.567: 0.990]
		SI	0.997 [0.971: 1.000]	0.915 [0.473: 0.995]	0.980 [0.693: 1.000]

Note. Values in brackets denote lower and upper quartiles.

The categorization of participants according to best-fitting model appears in Table 3.[16] Across both formats there is inconclusive support for the division of labor hypothesis, the percentage of individuals best described by an exemplar memory was somewhat higher in the non-additive than in the additive tasks (66% vs. 45%: $BF_{10} = 1.477$; $n = 76$). We replicate the results from previous studies in the sense that there is a shift from *additive* cue abstraction towards more exemplar memory in the non-additive tasks, and especially with the numerical format (i.e., from 14 vs. 5 to 0 vs. 12). However, in contrast to our predictions and previous studies there was as a non-trivial minority of participants (13, 33%) that appeared to engage in non-additive cue abstraction. The low SI for the numerical non-additive task in Table 2 indicates poor fit for

[16] Note the reported Bayesian contingency table tests compare number of participants categorized as relying on EBM with number of participants categorized as relying on CAM (additive and non-additive CAM collapsed). Thus participants with best support for the null model or uncategorized participants are excluded. They are though reported in Table 3 for transparency.

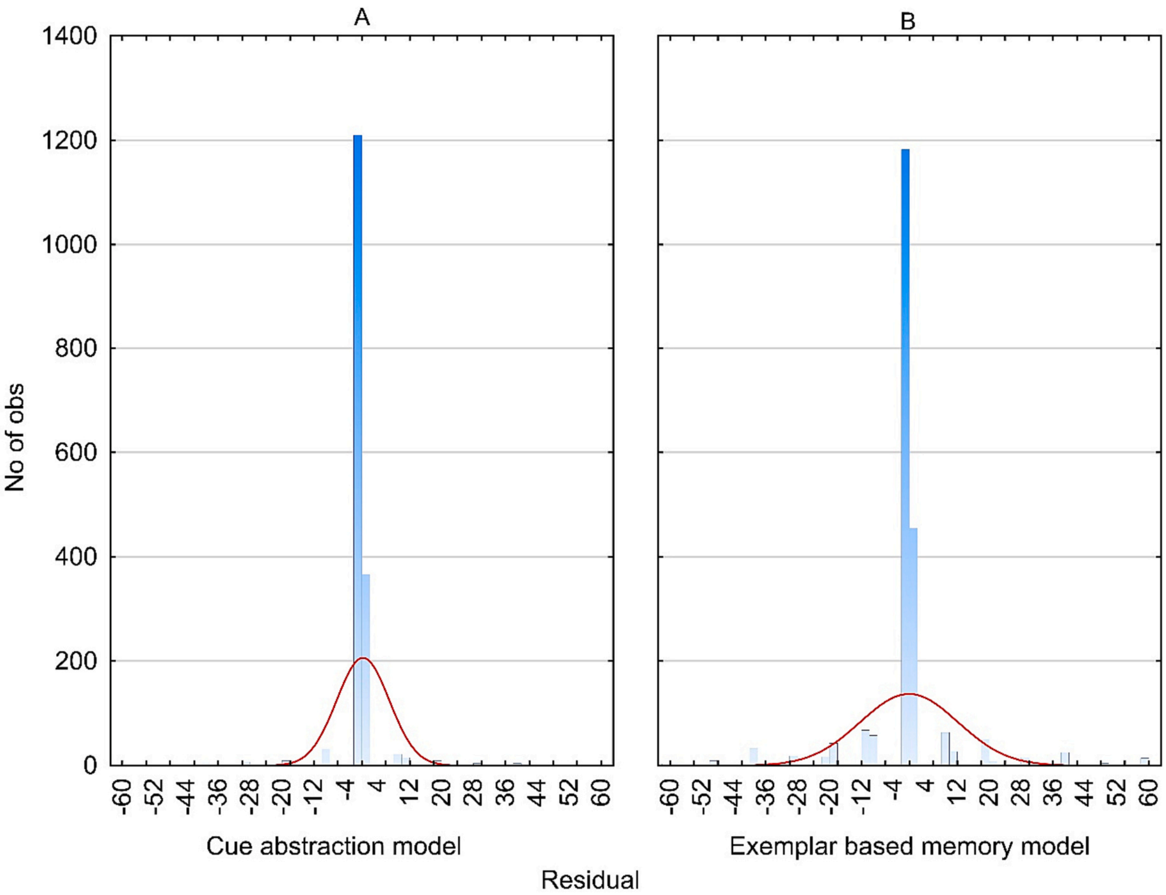


Fig. 4. Residual Distributions.
Note. The residuals from the predictions by the cue abstraction model (left) and the exemplar model (right) are distinctly leptokurtic and deviating from a normal distribution (in red), implying analytic processes with the PNP model.

some participants, but these participants are mainly those best fit by the exemplar based model, rather than those best fit by the non-additive cue abstraction model.

The rate of exemplar memory processes with the verbal and numerical format was similar (64% vs. 46%: $BF_{10} = 0.960$; $n = 76$). There was increased reliance on exemplar memory in the verbal-additive cell (63%) as compared to the numeric-additive cell (26%) ($BF_{10} = 4.789$, $n = 38$). Exemplar memory dominated both in the verbal non-additive task (65%) and in the numeric non-additive task (67%) ($BF_{10} = 0.368$, $BF_{01} = 2.717$, $N = 38$). However, while the percentage of individuals best described by an exemplar memory was higher in the numeric non-additive cell (67%) than the numeric additive cell (26%) ($BF_{10} = 7.368$ $n = 37$), exemplar memory dominated in both the verbal additive (63%) and the verbal non-additive cell (65%) ($BF_{10} = 0.368$, $BF_{01} = 2.717$, $n = 39$). The division of labor observed in previous studies was only observed with the numeric format, not with the verbal format.

2.3. Discussion

The results of Experiment 1 confirmed that a verbal magnitude format impedes learning in the additive task, but enhances learning in the non-additive task, as predicted if the verbal magnitude format decreases the initial rule bias that is advantageous in the additive task, but disadvantageous in the non-additive task. As in most previous studies, after training, in the test phase, with the numerical format (additive) rule-based cue abstraction with extrapolation was observed in the additive task, while exemplar memory, implying limited extrapolation, was the modal process in the non-additive task. By contrast, with the verbal magnitude format exemplar-based memory was the modal process both in the additive and the non-additive task. Note that while there was no difference in the number of participants in the non-additive tasks that had adopted exemplar memory in the test phase, the difference in performance suggests, in line with the prediction, that the participants with a verbal format adopted a memory strategy earlier, thus being able to fine-tune it and memorize items better. Model parameters (see Table 1C in Appendix C) also support such a conclusion with higher error-parameters in the numeric non-additive task than in the verbal non-additive task.

There were, however, also two important discrepancies between the results of Experiment 1 and the results from previous studies (e.g., Hoffmann et al., 2014, 2016; Juslin et al., 2003; Juslin et al., 2008; Karlsson et al., 2007; von Helversen & Rieskamp, 2009). First, there was a surprisingly high rate of non-additive cue abstraction. In the non-additive tasks, 13 participants were best described by the model that assumes that they abstract the cues and explicitly integrate them according to the non-additive equation (Eq. 2), allowing them to extrapolate also in the non-additive tasks. Such non-additive cue abstraction has rarely, if ever, been observed in the previous studies. However, as we discuss further in the General Discussion below, we find it unlikely that the participants have abstracted the exact algebraic Eq. 2 above and, literally speaking, number-crunch the cues according to Eq. 2. We rather suspect that they capture the non-additive relationship through the use of heuristic and sequential subspace strategies, which emulate the non-additive cue integration.

Second, in contrast to the assumption made in previous studies on multiple-cue judgment, the PNP modeling of the results from Experiment 1 suggested cognitive processes that typically came to exact deterministic outputs, occasionally marred by errors, rather than the

Table 3
Compilation of the Number of Participants Best Fitted by each Model, for each Cell and the Main Effects of Experiment 1.

			Task		
		Model	Additive	Non-additive	Main effect (Format)
Format	Numeric	CAM(A)	**14**	0	14
		CAM (NA)	0	6	6
		EBM	5	**12 (7 NC)**	**17 (7NC)**
		Uncateg.	0	1	1
		Null model	1	1	2
	Verbal	CAM(A)	7	0	7
		CAM (NA)	0	7	7
		EBM	**12**	**13 (7 NC)**	**25 (7 NC)**
		Uncateg.	0	0	0
		Null model	1	0	1
Main effect (Task)		CAM(A)	**21**	0	21
		CAM (NA)	0	13	13
		EBM	17	**25 (14 NC)**	**42 (14 NC)**
		Uncateg.	0	1	1
		Null model	2	1	3

Note: The modal model in each condition is denoted in bold font. CAM(A) refers to an additive cue abstraction model; CAM(NA) refers to a non-additive cue abstraction model; EBM to an exemplar-based model with either configural or non-configural coding; the null-model are participants best described by assuming that they always respond with their average response or the same response. Participants for whom BIC difference between the two best models were > -2 are marked as uncategorized. The modal model is marked in bold."NC" refers to exemplar models with non-configural coding that ignores the order of the two cues (see main text).

ubiquitously noisy output typical of intuitive cognitive processes (Sundh et al., 2021). In multiple-cue judgment, the cue integration in cue abstraction processes has typically been characterized as "quasi-rational", plagued by the inconsistencies associated with intuitive cue integration (see Hammond & Stewart, 2001; Karelaia & Hogarth, 2008). Likewise, exemplar memory has been assumed to involve intuitive retrieval processes (Nosofsky, 2011). These results suggest that participants best fit by a cue abstraction model seem to engage in number crunching an exact formula, and the participants best fit by an exemplar-based model to engage in rote-memorization. We return to this issue in the General Discussion.

3. Experiment 2: What numeric information elicits rule Bias

As is evident in Table 3, it is only in the numeric additive task, with *both* numeric cues and a numeric criterion, that cue abstraction is the modal model (app. 70%), in all other cells, exemplar memory is more frequent (app. 65%). The results of Experiment 1 did, however, not reveal if numerical cues alone are sufficient, if a numerical criterion alone is sufficient, or if both are necessary to elicit the initial cue abstraction that drives the division of labor observed with the numerical format, where cue abstraction dominates in the additive task but exemplar memory dominates in the non-additive task. In Experiment 2, we thus complemented the *congruent conditions* investigated in Experiment 1, where either both cues and criterion were numerical or verbal, with the corresponding *incongruent conditions*: either verbal cues and a numerical criterion or numerical cues and a verbal criterion.

Specifically, Experiment 2 allows us to contrast two alternative hypotheses. An *associative hypothesis* claims that (any) numeric information in the task changes participants expectations about the task, because the numbers elicit mathematical associations, such as that the properties represented are cardinal in nature and are related by some simple equation that can be identified, where the "simplicity" is suggestive of linear additive relations (see e.g., De Bock et al., 2002; De Bock et al., 2007; Dewolf et al., 2011; Van Dooren et al., 2008; Verschaffel et al., 1994). If this hypothesis is correct, introduction of numbers *either* in the cues or in the criterion should be *sufficient* to elicit the rule bias observed for the numerical formats in Experiment 1. This implies that we should observe the classical division of labor in all cells of Experiment 2, with cue abstraction in additive tasks and exemplar memory in non-additive tasks. On this hypothesis, both verbal cues and a verbal criterion are needed to inhibit the expectations of simple mathematical rules that hampers learning in the non-additive task in Experiment 1.

The *computational hypothesis* claims that it is crucial that both the cues and the criterion are numerical, because this allows the direct application of mathematical operations to the task content (as is the case in classical mathematical tasks where over-reliance on linearity is common, see e.g., Van Dooren et al., 2008; Dewolf et al., 2011; De Bock et al., 2007; De Bock et al., 2002; Verschaffel et al., 1994). If this hypothesis is correct both numerical cues and a numerical criterion is *necessary* to elicit the rule bias, and we should not observe a division-of-labor in any of the cells of Experiment 2, because none of the tasks in Experiment 2 have numbers both in the cues and the criterion, which allow direct computation based on the task contents. On this hypothesis, the rule bias only occurs if the task directly allows computation.[17]

In a later part of the Results section of Experiment 2 we benefit from the fact that Experiments 1 and 2 involve the same participant-population and jointly instantiate a complete factorial design that allows estimation of all main effects and interactions when independently manipulating cue and criterion formats in a complete factorial design.

3.1. Method

3.1.1. Participants

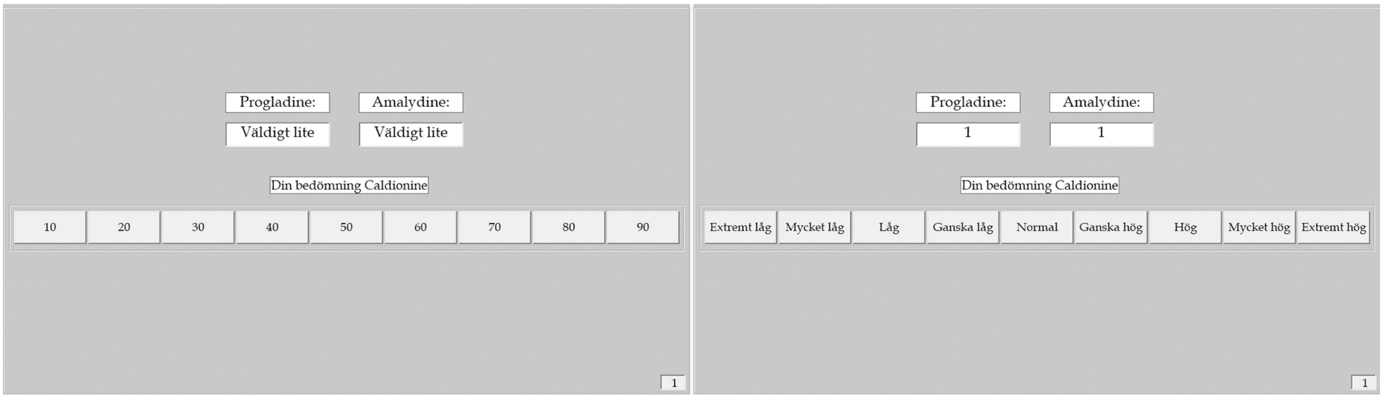
Eighty participants[18] (59 females, 20 males and 1 non-binary individual) ranging in age from 18 to 68 ($M = 26.86$, $SD = 7.96$) were recruited through public advertisement at various places at Uppsala University. Compensation was awarded in the form of a cinema voucher or (for students at the Department of Psychology) course credit.

3.1.2. Design

The experiment had a 2 × 2 between-subjects factorial design with format (verbal cues and numeric criterion or the reversed) and task (additive or non-additive) as independent between-subject variables. The dependent measure was the participants' judgments of the criterion and the accuracy of the judgments as measured by the RMSE between

[17] For reasons of transparency, we acknowledge that, while Experiment 2 tested the sufficiency conditions that are articulated in these two hypotheses already in the original versions of the article, the terms "associative hypothesis" and "computational hypothesis" were introduced in a later draft to simplify the exposition.

[18] Note that four participants were removed as outliers for analysis of performance during the test phase and three participants were removed as outliers for analysis of performance during the training-phase due to deviating >1.5 interquartile ranges from Q3 for RMSE (i.e. for having extremely high RMSE)



Progladine: Väldigt lite

Amalydine: Väldigt lite

Din bedömning Caldionine

Progladine: 1

Amalydine: 1

Din bedömning Caldionine

Fig. 5. Task Appearance for Verbal Cues (Left) and Numeric Cues (Right).
Note: The cue values are presented in the boxes with the headers “Progladine” and “Amalydine”. Below this it reads “Your judgment of Caldionine” and then follows the 9-step response scale.

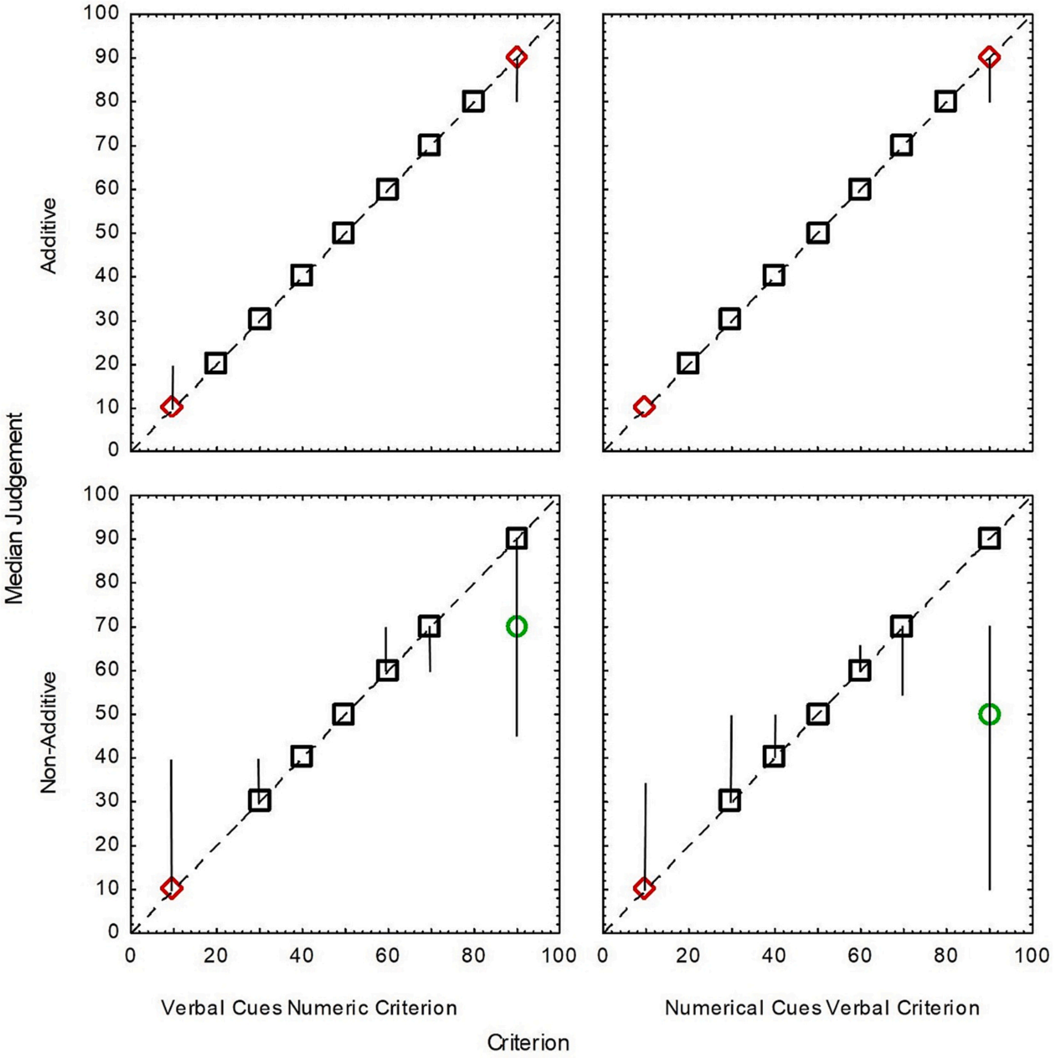


Fig. 6. Median Judgments in Experiment 2.
Note. The median judgments for each stimulus in the Test Phase as a function of the criterion value in each of the four cells of Experiment 2. The identity line represents perfectly accurate judgments. Diamond (red) items are extrapolation items given a configural exemplar-based memory strategy. Circle (green) items are extrapolation items for both a configural and a non-configural exemplar-based memory strategy in the non-additive tasks. Solid lines indicate interquartile ranges. Note that multiple points and lines overlap.

Table 4
Summary of Support for each Factor of a 2x2x10 Mixed Factorial BANOVA with Training Phase RMSE as the Dependent Variable.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Block	0.263	0.263	1.723e − 7	4.627e − 84	3.724e + 76
Cue-Format	0.263	0.263	0.420	0.175	2.394
Task	0.263	0.263	1.074e − 7	3.677e − 17	2.922e + 9
Block ✻ Cue-Format	0.263	0.263	0.001	0.824	0.001
Block ✻ Task	0.263	0.263	1.000	1.725e − 7	5.796e + 6
Cue-Format ✻ Task	0.263	0.263	0.404	0.420	0.962
Block ✻ Cue-Format ✻ Task	0.053	0.053	2.974e − 5	5.654e − 4	0.053

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF_{incl}) is the primary factor of concern showing the evidence for including a factor in the final model.

the judgment and the criterion.

3.1.3. *Material*

The material was identical to the material used in Experiment 1 with the exception that participants were either presented with verbal cues and a numeric criterion or numeric cues and a verbal criterion, as compared to Experiment1 where the cue-criterion format was congruent (see Fig. 5 for an example of how the task looked). The computer program used for stimulus presentation by default saved response times. The normative functions for inferring the criterion values were the same as in Experiment 1.

3.1.4. *Procedure & cognitive modeling*

The procedure was identical to Experiment 1, see 2.1.4. Procedure. Cognitive modeling was carried out in the same way as in Experiment 1, see 2.1.5. Cognitive modeling: The PNP model.

3.2. *Results and discussion*

Fig. 6 illustrates the median test phase judgments in each experimental cell of Experiment 2. These graphs indicate cue abstraction with accurate extrapolation in both of the additive cells. These results therefore suggest that having either numerical cues *or* a numerical criterion is sufficient to elicit initial cue abstraction that is successful in the additive task and that the median pattern indicating exemplar memory in the additive task occur only when both cues *and* criterion are verbal (as in Experiment 1, see Fig. 2A). The lower panels in Fig. 6 for the non-additive cells are similar to those for the corresponding cells of Experiment 1 (Fig. 2C-D), demonstrating inability to correctly judge new items in the higher criterion regions, but an ability to correctly judge the item in the lower regions.

3.2.1. *Performance in training*

A 2×2×10 mixed factorial BANOVA with RMSE as dependent variable and format (verbal cues and numeric criterion vs. numeric cues and verbal criterion) and task (additive vs. non-additive) as between-subject independent variables and training block (1–10) as the within-subject variable was computed. The two best supported but mutually indistinguishable models have a main effect of block, task, and format and an interaction of block and task ($BF_M = 13.01$, $BF_{10} > 10^{92}$, $BF_{2nd\ best\ model} = 1.039$) while the second model also includes an interaction between task and format.[19] All factors from the model with the strongest support are supported also when looking at their inclusion individually in Table 4.

The support for block and Block x Task is less interesting, as this is a training experiment where the participants, by design, learn from feedback to master two tasks that differ in speed of learning. The support for a main effect of task is the standard finding that people find it more difficult to learn non-additive than additive tasks (BF_{incl}. > 1,000,000,000: $M = 15.383$, $SD = 4.453$ for the non-additive task and $M = 6.814$, $SD = 4.676$ for the additive task). There is weak support (BF_{incl}. = 2.394) for better performance with numerical cues and a verbal criterion ($M = 10.129$, $SD = 6.396$) than with verbal cues and a numerical criterion ($M = 12.373$, $SD = 5.985$). As illustrated in Fig. 7, however, this effect is mainly driven by a difference in the learning performance in the additive task (numeric cues & verbal criterion $M = 4.716$, $SD = 3.687$ vs. verbal cues & numeric criterion $M = 8.801$, $SD = 4.725$), although this interaction is too weak to overcome the effect of a conservative prior (BF_{incl}. = 0.962).

As we will see in a later section (3.2.3. Cognitive modeling), additive cue abstraction dominates in the additive task and application of mathematical processing may be easier with numerical cues and a verbal criterion that only requires translation of one word into a number, than with verbal cues and a numerical criterion that requires two such translations of words into numbers. Maciejovsky and Budescu (2013) indeed show that incompatible formats lead to slower and worse integration due to translation. If the cue abstraction processes that dominate in the additive task are executed by numerical calculations that require such translation, whereas the exemplar memory processes in the non-additive task involve memory processes that require no verbal to number translations, this could explain why the format effects are mainly observed in the additive task.

Notably, in the non-additive task, performance is very poor regardless of the cue-criterion format combination (numeric cues & verbal criterion: $M = 15.000$, $SD = 3.844$, verbal cues & numerical criterion: $M = 15.767$, $SD = 5.061$: $BF_{10} = 0.347$; $BF_{01} = 2.884$). In sum: in Experiment 2, with numerical information in all the cells of the design, we only observed the standard finding that learning is slower with non-additive than with additive tasks, suggesting that the rule bias is operative in all of the cells, as implied by the associative hypothesis. This preliminary conclusions will be tested with cognitive modeling under 3.2.3 Cognitive modeling.

3.2.2. *Performance at test*

For the performance in the test phase, the model with the strongest support only includes a main effect of task ($BF_M = 7.922$, $BF_{10} > 10^7$, $BF_{2nd\ best\ model} = 3.747$),[20] with better performance in the additive task ($M = 3.416$, $SD = 3.756$) than in the non-additive task ($M = 13.499$, $SD = 7.424$). In contrast to Experiment 1, there is no interaction with format. Bayesian one-sample *t*-tests provide evidence against a difference in the performance between the (old) training items and the new extrapolation items in the additive task for both verbal cues and a numerical criterion (*Mean difference in RMSE* = 0.695, $SD = 5.464$, $n = 18$, $BF_{10} = 0.277$, $BF_{01} = 3.613$) and numerical cues and a verbal criterion ($M = 1.215$, $SD = 3.775$, $n = 18$, $BF_{10} = 0.538$, $BF_{01} = 1.857$), suggesting extrapolation and reliance on cue abstraction in the additive cells. Numerical cues or a numerical criterion, alone, appear sufficient to elicit cue abstraction, and since this is a successful strategy in the additive tasks, it is maintained into the test phase. The mean differences between training items and new items in both the verbal cues non-additive task

[19] See Appendix D for a full presentation of model comparison results.

[20] See Appendix D for a full presentation of model comparison results.

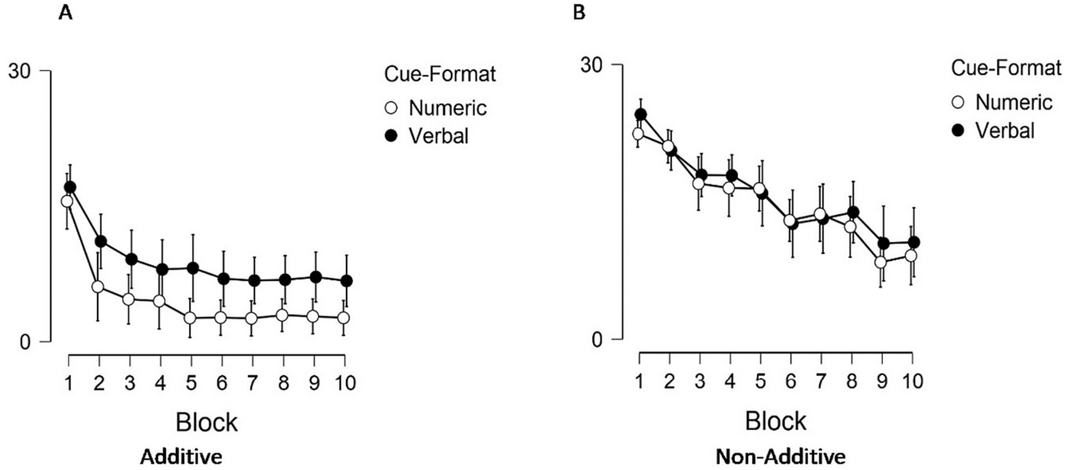


Fig. 7. Effects of Cue-Criterion Format and Cue-Criterion Relationship on Performance (RMSE) during Training.
Note. Effects of cue-format, cue-criterion relationship and training block on performance during training in the additive task (Panel A) and the non-additive task (Panel B). Error bars are 95% credible intervals.

Table 5
Median Adjusted R^2 and SI (Saturation Index) with Interquartile Index for the Best Fitting Model (Determined by BIC) for Each Individual Grouped by the Conditions in Experiment 2.

			Task		
		Index	Additive	Non-additive	Main effect (Format)
Format	Numeric cues, verbal criterion	R^2	0.990 [0.934: 1.000]	0.653 [0.309: 0.793]	0.839 [0.479: 0.992]
		SI	0.999 [0.994: 1.000]	0.679 [0.509: 0.929]	0.967 [0.604: 1.000]
	Verbal cues, numeric criterion	R^2	0.979 [0.844: 1.000]	0.602 [0.420: 0.984]	0.877 [0.452: 0.995]
		SI	0.997 [0.944: 1.000]	0.921 [0.542: 0.999]	0.986 [0.799: 1.000]
Main effect (Task)		R^2	0.989 [0.877: 1.000]	0.645 [0.383: 0.865]	0.877 [0.472: 0.995]
		SI	0.998 [0.963: 1.000]	0.783 [0.518: 0.987]	0.985 [0.649: 1.000]

Note. Values in brackets denote lower and upper quartiles.

Table 6
Compilation of the Number of Participants Best Fitted by each Model, for each Cell and the Main Effects of Experiment 2.

			Task		
		Model	Additive	Non-additive	Main effect (Format)
Format	Numeric cues, verbal criterion	CAM(A)	**13**	0	13
		CAM (NA)	0	1	1
		EBM	7	**18 (13 NC)**	**25(13NC)**
		Uncateg.	0	1	1
		Null model	0	0	0
	Verbal cues, numeric criterion	CAM(A)	**11**	1	12
		CAM (NA)	0	5	5
		EBM	7	**14 (7 NC)**	**21(7NC)**
		Uncateg.	0	0	0
		Null model	2	0	2
Main effect (Task)		CAM(A)	**24**	1	25
		CAM (NA)	0	6	6
		EBM	14	**32 (20 NC)**	**46 (20NC)**
		Uncateg.	0	1	1
		Null model	2	0	2

Note: The modal model in each condition is denoted in bold font. CAM(A) refers to an additive cue abstraction model; CAM(NA) refers to a non-additive cue abstraction model; EBM to an exemplar-based model with either configural or non-configural coding; the null-model are participants best described by assuming that they always respond with their average response or the same response. Participants for whom BIC difference between the two best models were > -2 are marked as uncategorized. The modal model is marked in bold.”NC” refers to exemplar models with non-configural coding that ignores the order of the two cues (see main text).

($M = 18.851$, $SD = 15.796$, $n = 20$, $BF_{10} = 666.772$) and the numerical cues non-additive task ($M = 32.197$, $SD = 16.944$, $n = 20$, $BF_{10} = 213{,}574.796$) were different from 0 with strong evidence, suggesting inability to extrapolate in the upper criterion regions. At the end of the training in the non-additive task, many participants have shifted to exemplar memory or some other process that constrains their ability to assess the new items.

3.2.3. Cognitive modeling

Cognitive modeling was carried out in the same way as in Experiment 1. The SI and Adj R^2 reported in Table 5 suggest that for all conditions except the numerical cues non-additive task, the best-fitting models account for most of the systematic variance in data. As in Experiment 1, for the numerical cues non-additive task, the best fitting model only accounts for app. two-thirds of the systematic variance in the data, suggesting that here the true cognitive process of all participants is not well captured by the models.[21]

The median parameter estimates for the coefficients of the cue abstraction model (Eqs. 4 & 5, see Appendix E) coincide exactly with the constants in the tasks (50, 10, 10 and 50, 10, 3 respectively). Median λ is close to zero in all conditions except the non-additive task with numeric cues and a verbal criterion, suggesting that the responses draw on analytic execution of the equations. Median λ was higher with exemplar memory than with cue abstraction (0.2 vs. 0.02; $BF_{10} = 1544.805$, $N = 77$, Bayesian Mann-Whitney test) and when the task was non-additive as compared to additive (0.2 vs. 0.02; $BF_{10} = 8.949$, $N = 77$, Bayesian Mann-Whitney test).

As is evident in Table 6, we have the division of labor in all format conditions, with additive cue abstraction in the additive tasks and exemplar memory in the non-additive tasks.[22] The percentage of individuals best described by an exemplar memory process was higher in the non-additive tasks than in the additive tasks (82% vs. 37%: $BF_{10} = 1078.712$; $n = 77$). The pattern is similar in both of the incongruent format conditions with very strong support in the numerical cues and verbal criterion condition (non-additive task 95% vs. additive task 35%: $BF_{10} = 975.381$, $n = 39$) and weaker support in the verbal cues and numeric criterion condition (non-additive task 70% vs. additive task 39%: $BF_{10} = 2.281$, $n = 38$). This again suggests support for the associate hypotheses, that as long as there is some numerical information in the task, participants start with cue abstraction processes, which – when successful – persist into the test phase.

As in Experiment 1, we observe a high rate of best fit for the non-additive cue abstraction model and, in contrast to the assumption made in much of the previous literature, most of the participants adopted analytic cognitive processes. The participants best fit by cue abstraction accordingly seem primarily to engage in number crunching of a formula and the participants best fit by an exemplar model seem to engage in rote-memorization.

3.2.4. Collapsed analyses

Experiment 1 and 2 can be viewed as cells in larger design with independent variables cue format (verbal/numerical), criterion format (verbal/numerical) and task (additive/non-additive). Because the experiments include non-overlapping samples of participants sampled from the same subject pool, we collapsed the two datasets into one in order to investigate interactions spanning the two experiments, as well as to increase the statistical power.

We performed a 2×2×2×10 mixed factorial BANOVA with cue format (verbal/numerical), criterion format (verbal/numerical), task (additive/non-additive) as between-subject independent variables and block (1–10) as the within-subject variable and RMSE as the dependent variable. The model with the strongest support includes a main effect of block, cue-format and task and an interaction effect of Block × Task and Cue-Format × Task ($BF_M = 151.773$, $BF_{10} > 10^{201}$, $BF_{2nd\ best\ model} = 1.836$).[23] All factors in the model, except the main effect of cue-format, are supported when looking at the factors individually.

There is inconclusive evidence against a main effect of cue-format ($BF_{incl} = 0.749$). Support for a two-way interaction between cue-format and task is substantial ($BF_{incl} = 26.316$). Additionally, there is weak support for a three-way interaction between block, task and cue-format when looking at the factors individually ($BF_{incl} = 2.113$).[24] These interactions are described in Fig. 8 below. As is evident from the graph the two-way and three-way interactions are driven by faster learning in the numerical cues conditions in the additive task ($M = 5.156$, $SD = 3.104$) in relation to the verbal cues conditions in the additive task ($M = 8.777$, $SD = 4.410$). In the non-additive task, average performance during training is similar between the numeric cues ($M = 15.943$, $SD = 3.661$) and the verbal cues ($M = 15.090$, $SD = 4.968$) conditions.

Fig. 9 compiles the proportions of participants best-fitted by the exemplar model in Experiment 1 and 2 as a function of the cue and criterion formats, with BF-factors from the Bayesian contingency tests previously presented in the result sections for each experiment. While there is a difference in the rate of exemplar memory in all cells with, at least some, numeric information, there is no difference when all information is presented in a verbal format. Again, this indicates that as long as some numerical information is present participants are invited to search for linear rules, which are successfully maintained into the test phase in the additive tasks.

4. Experiment 3: Replicating the beneficial effects of a verbal format in a non-additive environment

Experiment 2 suggests that the rule bias that slows down learning in the non-additive tasks was elicited in all cells of Experiment 2, as rule-based cue abstraction was the modal strategy in both of the additive cells. Because the presence of both verbal cues and a verbal criterion seems necessary to inhibit the rule bias that hampers learning in non-additive tasks, the beneficial effect of a verbal format in non-additive tasks in only tested in one cell across the two experiments (the cell with verbal cues and a verbal criterion in Experiment 1, all other cells included either numeric cues, a numeric criterion, or both). Because the beneficial effect of the verbal format in the non-additive task is a counter-intuitive, but important prediction, we wanted to replicate this result with an all-verbal format in a separate data collection.

4.1. Method

4.1.1. Participants

Ninety-nine[25] participants (32 females, 66 males and 1 non-binary individual) ranging in age from 21 to 71 ($M = 37.36$, $SD = 10.44$) located in the United Kingdom or the United States were recruited through Amazon Mechanical Turk. Participants received 7$ for conducting the experiment.

[21] Note that the SI distribution is positively skewed and the best fitting model accounts for almost all of the variance for six participants.

[22] Note the reported Bayesian contingency table tests compare number of participants categorized as relying on EBM with number of participants categorized as relying on CAM (additive and non-additive CAM collapsed). Thus participants with best support for the null model or uncategorized participants are excluded. They are though reported in Table 6 for transparency.

[23] See Appendix F, Table 1F for model comparison of the best 20 models.

[24] See Appendix F, Table 2F, for a full presentation of model average support for individual model factors.

[25] 100 participants were originally recruited, but one was excluded due to failing to answer questions separating an actual active human subject from a bot.

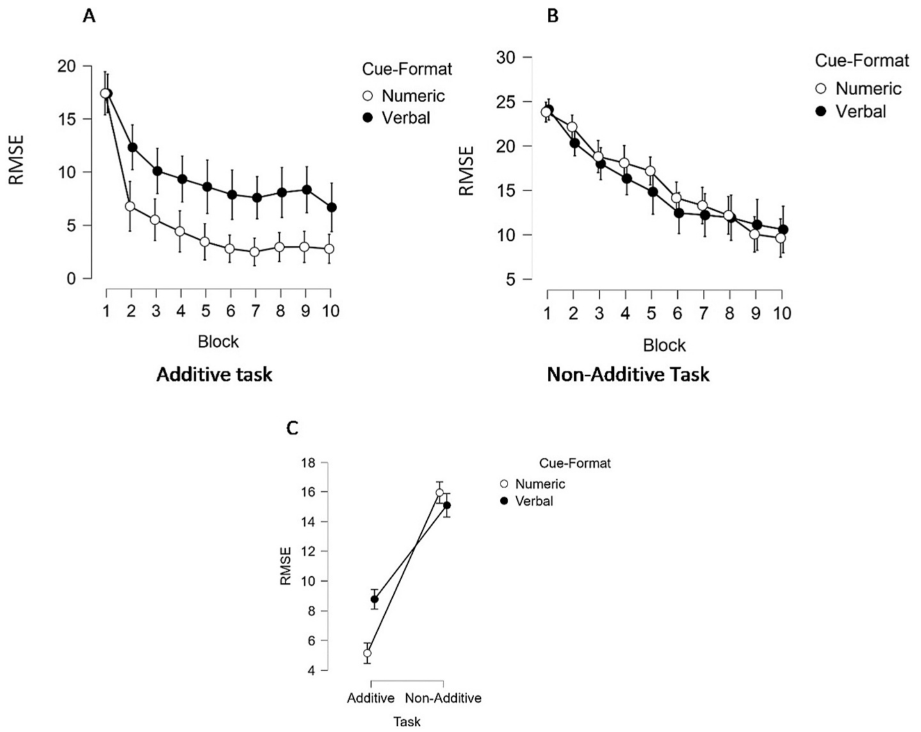


Fig. 8. Effects of Cue Format and Cue-Criterion Relationship on Performance (RMSE) during Training.

Note. Effects of cue format and cue-criterion relationship on performance (RMSE) during training for each block of training in the additive task (Panel A) and the non-additive task (Panel B). The effects of cue format and cue-criterion relationship on performance (RMSE) during training (Panel C). Error bars are 95% credible intervals. Data consists of all data from Exp1 and Exp2. Note that the scale of the Y-axis differs between panels.

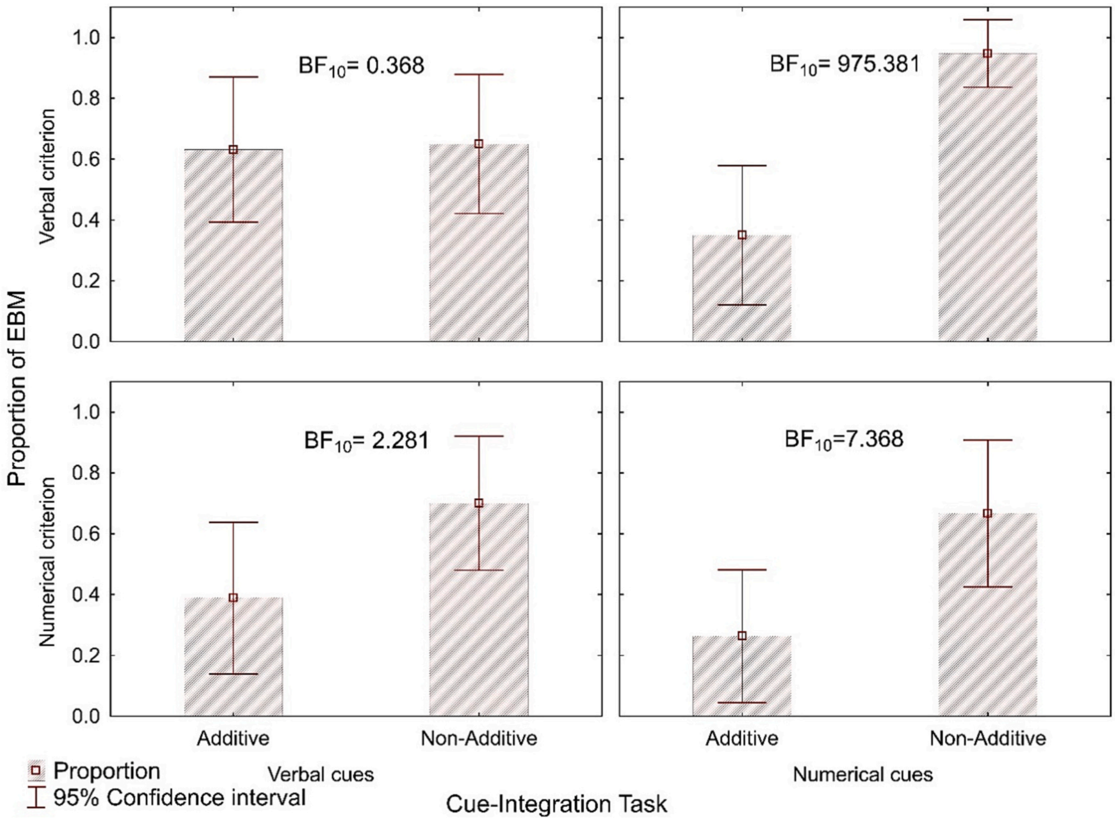


Fig. 9. Proportion of Participants Categorized as Relying on Exemplar-Based Memory (EBM) Depending on the Format of the Cues, the Format of the Criterion and the Task.

Note. The BF_{10} is the previously reported results from the Bayesian contingency table tests analyzing the difference in the proportion of participants relying on exemplar-based memory vs. cue abstraction in the additive and non-additive tasks in the respective experimental cells. Error bars are 95% Confidence intervals.

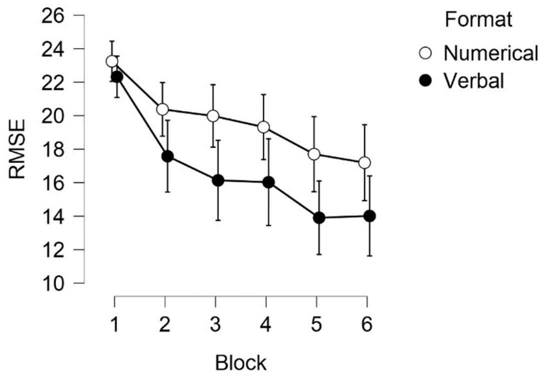


Fig. 10. Effects of Cue-Criterion Format on Performance (RMSE) during Training.

Note. Effects of Format (Verbal vs. Numerical) and Training Block (1–6) on performance during training. Participants are conducting a non-additive task. Error bars are 95% credible intervals.

4.1.2. *Design*

The experiment had a between-subjects design with format (verbal or numeric cues and criterion) as the independent between-subject variable. The dependent measure was the participants' judgments of the criterion and the accuracy of the judgments measured by the RMSE between the judgment and the criterion. All participants conducted the non-additive task from Experiment 1.

4.1.3. *Material*

The same material was used as in the non-additive task in Experiment 1, however the interface was altered slightly, presenting the trials with a survey program, and cue- and criterion labels translated to English, in order to facilitate online testing.

4.1.4. *Procedure*

Participants conducted the experiment via the online platform Amazon Mechanical Turk and were randomized to either of the two conditions. The training and test phase contained the same items as the non-**a**dditive task in Experiment 1, but the training was shortened to 6 blocks (rather than 10) to avoid fatigue effects with the AMT-format.

4.2. *Results and discussion*

A 2 × 6 mixed factorial BANOVA with RMSE as the dependent variable, format (verbal/numerical) as the between-subject independent variable and training block (1–6) as the within-subject variable showed that the best-supported model includes a main effect of format and block ($BF_M = 10.758$, $BF_{10} > 10^{27}$, $BF_{2nd\ best\ model} = 3.517$), with a $BF_{incl} = 3.516$ for a main effect of format.[26] The quicker learning of a non-additive task with a verbal ($M = 16.665$, $SD = 6.383$) rather than numerical ($M = 19.639$, $SD = 5.447$) format illustrated in Fig. 10 therefore replicates the result in Experiment 1 for an all-verbal format.

5. General discussion

The aim of this study was to investigate the effects of a verbal format of cues and criterion on the learning performance and the cognitive process adopted by the participants in a multiple-cue judgment task. The predictions for the study were based on two assumptions from the multiple-cue learning literature and a new hypothesis. The assumption of a *division of labor* between cognitive processes (Juslin et al., 2008) claims that while people can learn additive cue-criterion relations by controlled rule-based thinking (cue abstraction), they have to shift from cue abstraction to exemplar-based memory processes in tasks that require non-additive integration of the cue-criterion relations. The assumption of a *rule bias* (Ashby & Maddox, 2005; Juslin et al., 2008) claims that people typically start in a problem-solving mode when they learn categorization and multiple-cue judgment tasks, where they actively try to induce the underling structure of the task to arrive at explicit cue-criterion rules.

The new hypothesis was that a verbal format, in relation to a numerical one, should weaken this initial rule bias, facilitating a faster or immediate shift to exemplar memory in a non-additive task. Research suggests that through mathematical education we learn to apply linear mathematical models in mathematical contexts. While often helpful, this also leads to an over-application of linear rules in numerical contexts, impairing estimation in non-linear tasks (De Bock et al., 2002; De Bock et al., 2007; Dewolf et al., 2011; Ebersbach et al., 2008; Van Dooren et al., 2008; Verschaffel et al., 1994). We therefore hypothesized that mathematical training may establish an acquired association between numeric formats and use of linear, additive strategies, suggesting a potential format dependence of the results. It is thus plausible that the tendency to search for linear additive relations stems not only from cognitive constraints (Juslin et al., 2008) and linear additive models being prevalent and useful for prediction (Brehmer, 1994; Dawes & Corrigan, 1974), but also from mathematical education.

The two assumptions and the new hypothesis imply that a verbal format should impair learning in the additive environment (where cue abstraction is a viable strategy) but improve learning in the non-additive environment in relation to the numeric format (where a shift from cue abstraction to exemplar memory is needed). Further, we predicted that at the end of training most participants should have shifted to reliance on exemplar memory in the non-additive task, while many participants should still use cue abstraction in the additive task, especially in the numeric condition that more strongly invites cue abstraction. A division of labor should thus be observed with a numeric format, but not necessarily with a verbal format.

5.1. *Summary of the results*

In Experiment 1, we confirmed the predicted *interaction*: a verbal format *impaired* learning in the additive task but *enhanced* learning in the non-additive task – as we argued – by partially or wholly, de-activating the rule bias and the mathematical problem solving that is elicited initially by a numerical format. After training, in the numeric condition there was primarily cue abstraction in the additive task and exemplar memory in the non-additive task, but in the verbal condition, exemplar memory dominated in both tasks. This is in line with research suggesting that verbal formats invite more context dependent and associative reasoning (Liu et al., 2020a, 2020b; Windschitl & Wells, 1996). A numerical format is not always beneficial; under predictable conditions a verbal format allows faster learning.

In Experiment 2, we investigated what, and how much, numeric information that is needed to elicit the rule bias operative in the numerical condition of Experiment 1: Is a numeric format of both cues and criterion needed (so as to directly support a mathematical operation) to elicit the rule-bias that is a hindrance in learning a non-additive task, or is a numeric format of either cues or criterion sufficient to prime it. The results of Experiment 2 indicated that a numeric format of either the cues or the criterion is enough to elicit a rule-bias. This suggests that it is not the direct applicability of mathematical operations, as such, that is crucial, but rather that any numeric information changes the participant's expectations about the task, as implied if numeric formats invite higher expectations of a simple mathematical rule.

Benefitting from the fact that Experiments 1 and 2 involve the same participant population and jointly instantiate a complete factorial design that allows estimation of all main effects and interactions between cue- and criterion formats, aggregate analyses allowed us to address further questions. With regard to performance in training, the interaction effect

[26] See Appendix G for a full presentation of results.

of cue format and task seems the most important factor: cue format has the largest effect on speed of learning and drives the different learning rates in the additive and non-additive tasks. This suggests that a numerical format of the cues traps the participant in (successful or futile) attempts at cue abstraction for a longer time than a numerical format of the criterion. The results, however, suggest that complete absence of numeric information is needed for clear improvement in the non-additive task, as seen in Experiment 1 and the replication.

With regard to eliciting stronger reliance on cue abstraction in the test phase, an interaction is the most important effect (see Fig. 9): When both the cues and the criterion are verbal, exemplar memory dominates both in the additive and in the non-additive task, but as soon as there is any numerical information in the task, cue abstraction is more prevalent in the additive task and exemplar memory is more prevalent in the non-additive task. A numeric format in either cues or criterion is sufficient to invite more initial cue abstraction, which, however, only survives into the test phase in additive tasks, where it is successful. This reaffirms the conclusion suggested already by the analysis of Experiment 2 alone.

The key predicted interaction between format and task was observed both in Experiment 1 and the collapsed analysis of Experiment 1 and 2, but the most counter-intuitive part of this interaction – that a verbal format can actually enhance learning in a non-additive task – was only tested and confirmed by a comparison between two cells in Experiment 1 (i.e., where both cues and criterion have the same format). Thus, we performed Experiment 3 with the critical all-verbal condition that enhanced learning in the non-additive task and replicated the beneficial effect of a verbal format.

To conclude, we draw the conclusion that the tendency to shift from rule-based cue abstraction to exemplar-based memory processes as a function of task-properties, as shown in numerous studies, seems to be considerably larger in the presence of numeric formats. When all information is verbal participants appear more rapidly to home in on an exemplar-based memory strategy regardless of the task-properties. This provides an explanation for why the verbal format is beneficial for learning in the non-additive environment, namely that if participants engage exemplar-based memory early on, their learning will be faster and at the end of training the given exemplar-based memory process will be more fine-tuned.

5.2. *"Anomalies" relative to previous results*

We now turn to other discrepancies from previous results, namely: **i**) a surprising number of participants seem to adopt a non-additive integration strategy (13 in Experiment 1 and 6 in Experiment 2) and **ii**) most participants address the task in an analytic manner. In regard to the first result, we do not believe that the participants best fit by the non-additive cue abstraction model really have induced the rather complex nonlinear function (Eq. 2) and that they literally crunch the cues according to this equation. Rather, we hypothesize that they rely on sub-space strategies. They partition the stimulus space into subspaces and adopt a range of strategies that successfully emulate non-additive integration at the level of the whole space (see e.g., Kalish et al., 2004). This could for example entail either adopting multiple different linear additive rules depending on the cue values present (theoretically the non-additive task can be solved by adopting 5 different linear additive rules); identifying one or several linear additive rules that successfully capture the relations between cues and criterion in some areas of the stimulus space and memorizing the remaining cue-criterion combinations, or a sequential decision-tree which identifies a few contingencies that together emulate the non-additive judgments. This proposal is purely speculative at present and need to be examined in future research.

Concerning the second point, most participants solved the task by analytical exact application of an algorithm (Sundh et al., 2021). Thus, participants best fit by a cue abstraction model in this paper are not integrating the cues by an informal and inconsistent process (as suggested by e.g., Brehmer, 1994; Karelaia & Hogarth, 2008; Juslin et al., 2008), rather they number crunch an exact formula (either the normative rule, multiple rules or a formalized decision tree). Participants best fit by an exemplar-based modal similarly produce exact responses. This suggests that the participants are relying more on rote-memorization of exemplars, rather than on a similarity-based inference. Because previous studies have not drawn on the PNP model, this possibility has not been previously investigated.

Interestingly, we find no support for more analytical processes (e.g., number crunching rather than informal integration) when the format is numerical as compared to verbal. This in apparent contrast to findings suggesting that verbal magnitude formats are processed more intuitively than numerical ones (Liu et al., 2020b; Wallsten et al., 1993; Windschitl & Wells, 1996). Note, however, that this is because, at least in the test phase, the performance based on exemplar memory in these tasks take the form of overlearned rote-memorization of individual exemplars. In tasks where people cannot draw on rote memorizing each individual exemplar it may still be true that associative and similarity-based exemplar inference is primarily intuitive. The previous studies referred to above have also been performed in the context of the dual-systems framework (e.g., Evans, 2008), which need not overlap with the operational Brunswikian definition in the PNP model (Sundh et al., 2021).

5.3. *Limitations and future directions*

A first foundational limitation refers to the possibility to empirically distinguish between abstract rule-based representations and memories of concrete exemplars. As noted by Barsalou (1990), assumptions about representations can only be examined in the context of additional processing assumptions, so that what we can test and compare are always specific representation-process conjunctions, rather than general claims about representation.

This is what we do in this study when we compare cue abstraction and exemplar models. The conjunctions between representations and processes embodied in these models are, however, not arbitrary but tested (and frequently supported) in numerous studies (see the review in the Introduction). It is well-known that these two models can be empirically distinguished in data, for example, by the degree to which they allow extrapolation beyond the training range and by their ability to produce accurate judgments also in highly nonlinear environments. Our study is concerned with empirically identifying conditions under which we find more relative support for one rather than the other of these two models. But the results cannot be taken to motivate unqualified and open-ended universal claims about the role of abstract and concrete representations in any shape or form in these tasks.

A related limitation refers to the (plausible) possibility that judgments are based on mixes between rule-based and exemplar processes within the same task, participant or even trial (Bröder et al., 2017; Izydorczyk & Bröder, 2021). The present studies were not designed to test this possibility, which needs to be examined in future research. We, however, believe that our results can be convincingly interpreted also within such a mixed-process framework. These mixed-process models represent the observed judgments as a weighted combination of judgments produced by rule-based and exemplar-based cognitive processes. In this framework, our results translate into the conclusion that numbers seem to lead to judgments with more initial weight assigned to rule-based cue abstraction processes, whereas the verbal formats seem to invite judgments with a larger initial weight for exemplar processes.

Another limitation is that the experiments are conducted within a narrow experimental paradigm, and in order to strengthen the conclusions and the generalizability across contexts future research comparing numerical and verbal formats within other contexts, both in regards to for example the number of cues and the cue-criterion relationships (i.e., the exact algorithms that relate the cues to the criterion) are warranted. In addition, the manipulation of the formats themselves can be expanded to include, for example, longer texts as verbal information, and different measures both for performance, and for investigating the cognitive process can be used for further generalizability across methodologies. While we acknowledge this limitation, which plagues much of research in cognitive psychology, we believe that this issue of generalization needs treatment beyond what we can pursue in this article.

That some participants provide exactly correct responses in the non-additive environment raises the question of what type of process these participants are relying on. An interesting endeavor for future research is to design environments where sub-space strategies that rely on adopting multiple linear rules are either possible or not possible and to continue the work on the developing models to successfully capture cognitive processes that rely on both memory and cue abstraction (as, e.g., the CX-COM, see Albrecht et al., 2020).

The comparatively bad fit of the best fitting model in the numeric non-additive environment raises questions. What are participants in this condition doing? The informal inspection of individual data patterns suggest that they might be relying on a mixed strategy where the old-items have been rote-memorized. Participants may be memorizing items throughout training, but fall back on linear additive integration for new items in the test phase, or when facing new items retrieve an item probabilistically rather than retrieving the most similar item. As noted above, combined strategies, both multiple cue abstraction strategies and combinations of exemplar memory and cue abstraction are relevant avenues for future research (see Albrecht et al., 2020; Bröder et al., 2017; Izydorczyk & Bröder, 2021).

6. Conclusions

A verbal magnitude format is beneficial for learning in a non-additive multiple-cue environment, but detrimental for learning in an additive environment as compared to a numerical format. This is an effect of the cognitive process adopted. As long as any numeric information is present, participants are invited to actively search for linear additive rules, which is helpful in an additive task, but detrimental in a non-additive task. Conversely, with a verbal magnitude format, participants home in on reliance on exemplar-based memory. Thus, a division of labor between rule-based processes in additive tasks and exemplar-based processes in non-additive tasks may be contingent on the presence of numeric information. At a more paradigm-critical level, the results also illustrate how apparently trivial choices of convenience in the design of the experimental tasks, like the common use of numerical formats, can have strong substantive implications for the conclusions obtained.

Author note

The cognitive modeling of cognitive strategy and process from Experiment 1 and Experiment 2 are used in aggregate and summary analyses (together with data from other samples) in a book chapter, Collsiöö et al. (2023). The aggregate analyses in the book chapter regard different research questions than the ones in the submitted manuscript.

CRediT authorship contribution statement

August Collsiöö: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Peter Juslin:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Anders Winman:** Conceptualization, Software, Writing – review & editing.

Declaration of Competing Interest

None.

Data availability

The Raw- and processed data analyzed in the present paper, as well as MATLAB code for use of the PNP model, are available on OSF via the following link: https://osf.io/qx6gt/

Acknowledgements

The research was funded by the Swedish Research Council (Grant 2019-03121) and the Marcus and Amalia Wallenberg Foundation (MAW 2016.0132). The funding source had no role other than financial support. All listed authors contributed in a significant way to the manuscript and have read and approved the final manuscript. We are heavily indebted to Ebba Elwin for her thoughtful comments and contribution to the theoretical base of the project and to Sarah Ericsson for assistance with data collection.

Appendix A. List of numerical items

Table 1A
Items for the additive and non-additive numerical conditions.

Item	Progladine (cue1)	Amalydine (cue2)	Additive criterion (Caldionine)	Additive Included in training (1)	Non-additive criterion (Caldionine)	Non-additive Included in training (1)
1	1	1	50	1	90	0
2	1	2	40	1	70	1
3	1	3	30	1	50	1
4	1	4	20	1	30	1
5	1	5	10	0	10	1
6	2	1	60	1	70	1
7	2	2	50	1	60	1
8	2	3	40	1	50	1
9	2	4	30	1	40	1
10	2	5	20	1	30	1
11	3	1	70	1	50	1
12	3	2	60	1	50	1
13	3	3	50	1	50	1
14	3	4	40	1	50	1
15	3	5	30	1	50	1
16	4	1	80	1	30	1
17	4	2	70	1	40	1
18	4	3	60	1	50	1

(continued on next page)

Table 1A (*continued*)

Item	Progladine (cue1)	Amalydine (cue2)	Additive criterion (Caldionine)	Additive Included in training (1)	Non-additive criterion (Caldionine)	Non-additive Included in training (1)
19	4	4	50	1	60	1
20	4	5	40	1	70	1
21	5	1	90	0	10	0
22	5	2	80	1	30	1
23	5	3	70	1	50	1
24	5	4	60	1	70	1
25	5	5	50	1	90	1

Note. Items with a zero (0) in the “Included in training” column were omitted during training in order to differentiate if participants adopted an exemplar-based memory strategy or a rule-based strategy to solve the task.

Appendix B. Exp.1 Model comparison of factors influencing performance (RMSE)

Table 1B
Training-Phase Model Comparison Mixed Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF $_M$	BF $_{10}$	error %
Null model (incl. subject)	0.053	1.339e − 97	2.410e − 96	1.000	
Block + Task + Format + Block ✻ Task + Task ✻ Format	0.053	0.678	37.968	5.067e + 96	4.414
Block + Task + Block ✻ Task	0.053	0.237	5.583	1.769e + 96	1.067
Block + Task + Format + Block ✻ Task	0.053	0.053	1.008	3.960e + 95	1.817
Block + Task + Format + Block ✻ Task + Block ✻ Format + Task ✻ Format + Block ✻ Task ✻ Format	0.053	0.016	0.299	1.220e + 95	2.228
Block + Task + Format + Block ✻ Task + Block ✻ Format + Task ✻ Format	0.053	0.014	0.254	1.039e + 95	2.702
Block + Task + Format + Block ✻ Task + Block ✻ Format	0.053	0.001	0.020	8.476e + 93	3.584
Block + Task + Format + Task ✻ Format	0.053	3.237e −4	0.006	2.418e + 93	4.011
Block + Task	0.053	1.206e −4	0.002	9.008e + 92	1.493
Block + Task + Format	0.053	2.549e −5	4.589e −4	1.904e + 92	2.290
Block + Task + Format + Block ✻ Format + Task ✻ Format	0.053	7.307e −6	1.315e −4	5.458e + 91	4.064
Block + Task + Format + Block ✻ Format	0.053	5.408e −7	9.734e −6	4.039e + 90	1.743
Block	0.053	7.314e − 15	1.317e − 13	5.464e + 82	0.346
Block + Format	0.053	2.137e − 15	3.846e − 14	1.596e + 82	1.031
Block + Format + Block ✻ Format	0.053	4.483e − 17	8.069e − 16	3.348e + 80	0.981
Task + Format + Task ✻ Format	0.053	2.961e − 87	5.329e − 86	2.212e + 10	2.399
Task	0.053	1.769e − 87	3.185e − 86	1.322e + 10	1.352
Task + Format	0.053	2.830e − 88	5.093e − 87	2.114e + 9	2.048
Format	0.053	2.944e − 98	5.299e − 97	0.220	0.943

Note. All models include subject. The table starts with the Null Mode. All subsequent models are ordered from the one with the strongest support to the one the weakest support.

Table 2B
Training-Phase Analysis of the Effects of Included Factors.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Block	0.263	0.263	4.698e − 4	5.013e − 87	9.372e + 82
Task	0.263	0.263	1.466e − 4	9.496e − 15	1.544e + 10
Format	0.263	0.263	0.053	0.237	0.224
Block ✻ Task	0.263	0.263	0.983	4.777e −4	2058.392
Block ✻ Format	0.263	0.263	0.015	0.732	0.021
Task ✻ Format	0.263	0.263	0.693	0.054	12.786
Block ✻ Task ✻ Format	0.053	0.053	0.016	0.014	1.174

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF$_{incl}$) is the primary factor of concern showing the evidence for including a factor in the final model.

Table 3B
Test Phase Model Comparison Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF $_M$	BF $_{10}$	error %
Null model	0.200	1.552e − 5	6.206e − 5	1.000	
Task + Format + Task ✻ Format	0.200	0.454	3.323	29,243.628	1.555
Task	0.200	0.433	3.051	27,885.351	3.588e − 8
Task + Format	0.200	0.114	0.512	7319.683	1.679
Format	0.200	4.066e − 6	1.626e − 5	0.262	0.014

Note. The table starts with the Null Mode. All subsequent models are ordered from the one with the strongest support to the one the weakest support.

Table 4B
Test Phase Analysis of the Effects of Included Factors.

Analysis of Effects - RMSE					
Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Task	0.400	0.400	0.546	1.958e − 5	27,894.988
Format	0.400	0.400	0.114	0.433	0.262
Task ✻ Format	0.200	0.200	0.454	0.114	3.995

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF$_{incl}$) is the primary factor of concern showing the evidence for including a factor in the final model.

Appendix C. Exp. 1 Median parameters from cognitive model fitting

Table 1C
Compilation of Median Parameter Estimates (λ, α, ω β and σ) for Participants Best Fit by the Additive Cue Abstraction Model (CAM(A)), the Non-Additive Cue Abstraction Model (CAM(NA)), and the Exemplar-Based Model (EBM).

Condition	CAM(A)					CAM(NA)					EBM		
	λ	α	ω_1	ω_2	σ^*	λ	α	ω_1	ω_2	σ^*	λ	β	σ^*
Verbal-Additive	0.04	50	10	10	9.99(10.14)						0.16	17.00	18.87
Numeric-Additive	0	50	10	10	0 (20)						0.20	15.54	14.14
Verbal-Non-Additive						0.02	50	10	3	14.14(29.27)	0.24	16.27	22.36
Numeric-Non-Additive						0.06	50	10	3	22.69(26.97)	0.17	17.74	31.04

Note: The $\boldsymbol{\sigma}$ -value without parenthesis is the median error variance across all participants in each condition, where participants that only produced exact responses ($\lambda = 0$) have been assigned zero error variance. The values within parenthesis is the median error variance across the participants that occasionally produced responses with error variance ($\lambda > 0$). See Sundh et al. (2021) for a discussion of these different ways of reporting the error variance with the PNP model.

Appendix D. Exp. 2 Model comparison of factors influencing performance (RMSE)

Table 1D
Training-Phase Model Comparison Mixed Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF $_M$	BF $_{10}$	error %
Null model (incl. subject)	0.053	6.620e − 94	1.192e − 92	1.000	
Block + Cue-Format + Task + Block ✻ Task	0.053	0.420	13.014	6.339e + 92	2.639
Block + Cue-Format + Task + Block ✻ Task + Cue-Format ✻ Task	0.053	0.404	12.196	6.102e + 92	5.724
Block + Task + Block ✻ Task	0.053	0.175	3.825	2.648e + 92	1.057
Block + Cue-Format + Task + Block ✻ Cue-Format + Block ✻ Task	0.053	6.370e −4	0.011	9.624e + 89	2.236
Block + Cue-Format + Task + Block ✻ Cue-Format + Block ✻ Task + Cue-Format ✻ Task	0.053	5.654e −4	0.010	8.541e + 89	2.586
Block + Cue-Format + Task + Block ✻ Cue-Format + Block ✻ Task + Cue-Format ✻ Task + Block ✻ Cue-Format ✻ Task	0.053	2.974e −5	5.353e −4	4.493e + 88	2.231
Block + Cue-Format + Task	0.053	7.643e −8	1.376e −6	1.155e + 86	4.227
Block + Cue-Format + Task + Cue-Format ✻ Task	0.053	6.498e −8	1.170e −6	9.817e + 85	3.936
Block + Task	0.053	3.090e −8	5.562e −7	4.668e + 85	0.860
Block + Cue-Format + Task + Block ✻ Cue-Format	0.053	1.167e − 10	2.100e −9	1.762e + 83	5.718
Block + Cue-Format + Task + Block ✻ Cue-Format + Cue-Format ✻ Task	0.053	1.003e − 10	1.805e −9	1.515e + 83	3.465
Block	0.053	2.001e − 17	3.602e − 16	3.023e + 76	0.314
Block + Cue-Format	0.053	1.674e − 17	3.012e − 16	2.528e + 76	1.303
Block + Cue-Format + Block ✻ Cue-Format	0.053	2.524e − 20	4.543e − 19	3.813e + 73	1.378
Cue-Format + Task	0.053	2.064e − 84	3.715e − 83	3.118e + 9	1.877
Cue-Format + Task + Cue-Format ✻ Task	0.053	1.554e − 84	2.797e − 83	2.347e + 9	2.134
Task	0.053	1.009e − 84	1.816e − 83	1.524e + 9	0.729
Cue-Format	0.053	4.720e − 94	8.496e − 93	0.713	0.981

Note. All models include subject. The table starts with the Null Mode. All subsequent models are ordered from the one with the strongest support to the one the weakest support.

Table 2D
Training-Phase Analysis of the Effects of Included Factors.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Block	0.263	0.263	1.723e − 7	4.627e − 84	3.724e + 76
Cue-Format	0.263	0.263	0.420	0.175	2.394
Task	0.263	0.263	1.074e − 7	3.677e − 17	2.922e + 9
Block ✻ Cue-Format	0.263	0.263	0.001	0.824	0.001
Block ✻ Task	0.263	0.263	1.000	1.725e −7	5.796e + 6
Cue-Format ✻ Task	0.263	0.263	0.404	0.420	0.962
Block ✻ Cue-Format ✻ Task	0.053	0.053	2.974e − 5	5.654e −4	0.053

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF$_{\text{incl}}$) is the primary factor of concern showing the evidence for including a factor in the final model.

Table 3D
Test Phase Model Comparison Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF $_M$	BF $_{10}$	error %
Null model	0.200	1.998e − 8	7.991e − 8	1.000	
Task	0.200	0.664	7.922	3.326e + 7	1.709e − 11
Task + Cue-Format	0.200	0.177	0.862	8.876e + 6	1.507
Task + Cue-Format + Task ✻ Cue-Format	0.200	0.158	0.752	7.919e + 6	2.543
Cue-Format	0.200	5.191e − 9	2.076e − 8	0.260	0.017

Note. The table starts with the Null Mode. All subsequent models are ordered from the one with the strongest support to the one the weakest support.

Table 4D
Test Phase Analysis of the Effects of Included Factors.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Task	0.400	0.400	0.842	2.517e − 8	3.345e + 7
Cue-Format	0.400	0.400	0.177	0.664	0.267
Task ✻ Cue-Format	0.200	0.200	0.158	0.177	0.892

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF$_{\text{incl}}$) is the primary factor of concern showing the evidence for including a factor in the final model.

Appendix E. Exp. 2 Median parameters from cognitive model fitting

Table 1E
Compilation of Median Parameter Estimates (λ, α, ω, β and σ) for Participants Best Fit by the Additive Cue-Abstraction Model (CAM(A)), the Non-Additive Cue-Abstraction Model (CAM(NA)), and the Exemplar-Based Model (EBM).

Condition	CAM(A)					CAM(NA)					EBM		
	λ	α	ω_1	ω_2	σ^*	λ	α	ω_1	ω_2	σ^*	λ	β	σ^*
Verbal Cues Additive	0.02	50	10	10	9.06 (10)						0.18	15.64	14.90
Numeric Cues Additive	0.02	50	10	10	4.78 (9.14)						0.08	17.48	22.34
Verbal Cues Non-Additive	0.76	−10.1	10	−10	16.91	0	50	10	3	0(10)	0.26	16.95	26.55
Numeric Cues Non-Additive						0.30	50	10	3	47.81	0.21	18.29	29.11

Note: The σ -value without parenthesis is the median error variance across all participants in each condition, where participants that only produced exact responses (λ = 0) have been assigned zero error variance. The values within parenthesis is the median error variance across the participants that occasionally produced responses with error variance ($\lambda > 0$). See Sundh et al. (2021) for a discussion of these different ways of reporting the error variance with the PNP model.

Appendix F. Collapsed analysis Model comparison of factors influencing performance (RMSE)

Table 1F
Training-Phase Model Comparison Mixed Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF$_M$	BF$_{10}$	error %
Null model (incl. subject and random slopes)	0.006	1.205×10^{-202}	2.000×10^{-200}	1.000	
Block + Cue-Format + Task + Block✻ Task + Cue-Format✻ Task	0.006	0.478	151.773	$3.965 \times 10^{+201}$	2.167
Block + Cue-Format + Criterion-Format + Task + Block✻ Task + Cue-Format✻ Task	0.006	0.260	58.366	$2.160 \times 10^{+201}$	1.513
Block + Cue-Format + Criterion-Format + Task + Block✻ Task + Cue-Format✻ Task + Criterion-Format✻ Task	0.006	0.081	14.555	$6.692 \times 10^{+200}$	2.071
Block + Cue-Format + Criterion-Format + Task + Cue-Format✻ Criterion-Format + Block✻ Task + Cue-Format✻ Task	0.006	0.074	13.183	$6.108 \times 10^{+200}$	13.174
Block + Task + Block✻ Task	0.006	0.024	4.071	$1.987 \times 10^{+200}$	0.783

(*continued on next page*)

Table 1F (*continued*)

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Block + Cue-Format + Criterion-Format + Task + Cue-Format✻ Criterion-Format + Block✻ Task + Cue-Format✻ Task + Criterion-Format✻ Task	0.006	0.019	3.301	$1.618 \times 10^{+200}$	2.168
Block + Cue-Format + Task + Block✻ Task	0.006	0.018	3.080	$1.512 \times 10^{+200}$	1.116
Block + Criterion-Format + Task + Block✻ Task	0.006	0.014	2.278	$1.124 \times 10^{+200}$	0.978
Block + Cue-Format + Criterion-Format + Task + Block✻ Task	0.006	0.010	1.665	$8.244 \times 10^{+199}$	1.039
Block + Cue-Format + Criterion-Format + Task + Cue-Format✻ Criterion-Format + Block✻ Task + Cue-Format✻ Task + Criterion-Format✻ Task + Cue-Format✻ Criterion-Format✻ Task	0.006	0.006	0.968	$4.813 \times 10^{+199}$	2.430
Block + Criterion-Format + Task + Block✻ Task + Criterion-Format✻ Task	0.006	0.004	0.713	$3.548 \times 10^{+199}$	6.784
Block + Cue-Format + Criterion-Format + Task + Block✻ Task + Criterion-Format✻ Task	0.006	0.003	0.519	$2.586 \times 10^{+199}$	3.835
Block + Cue-Format + Criterion-Format + Task + Cue-Format✻ Criterion-Format + Block✻ Task	0.006	0.003	0.431	$2.150 \times 10^{+199}$	0.968
Block + Cue-Format + Task + Block✻ Cue-Format + Block✻ Task + Cue-Format✻ Task + Block✻ Cue-Format✻ Task	0.006	0.002	0.358	$1.788 \times 10^{+199}$	2.615
Block + Cue-Format + Criterion-Format + Task + Block✻ Cue-Format + Block✻ Task + Cue-Format✻ Task + Block✻ Cue-Format✻ Task	0.006	0.001	0.200	$1.000 \times 10^{+199}$	3.026
Block + Cue-Format + Task + Block✻ Cue-Format + Block✻ Task + Cue-Format✻ Task	0.006	0.001	0.179	$8.955 \times 10^{+198}$	9.126
Block + Cue-Format + Criterion-Format + Task + Cue-Format✻ Criterion-Format + Block✻ Task + Criterion-Format✻ Task	0.006	7.798×10^{-4}	0.130	$6.474 \times 10^{+198}$	1.396
Block + Cue-Format + Criterion-Format + Task + Block✻ Cue-Format + Block✻ Task + Cue-Format✻ Task	0.006	5.038×10^{-4}	0.084	$4.182 \times 10^{+198}$	1.256
Block + Cue-Format + Criterion-Format + Task + Block✻ Cue-Format + Block✻ Task + Cue-Format✻ Task + Criterion-Format✻ Task + Block✻ Cue-Format✻ Task	0.006	3.542×10^{-4}	0.059	$2.940 \times 10^{+198}$	2.232

Note. All models include subject, and random slopes for all repeated measures factors.
Note. Showing the best 20 out of 167 models.

Table 2F
Training-Phase Analysis of the Effects of Included Factors.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF_{incl}
Block	0.114	0.114	1.058×10^{-13}	1.383×10^{-180}	$7.649 \times 10^{+166}$
Cue-Format	0.114	0.114	0.031	0.042	0.749
Criterion-Format	0.114	0.114	0.285	0.523	0.546
Task	0.114	0.114	7.095×10^{-15}	9.840×10^{-36}	$7.210 \times 10^{+20}$
Cue-Format✻ Criterion-Format	0.299	0.299	0.097	0.356	0.272
Cue-Format✻ Task	0.299	0.299	0.914	0.035	26.316
Criterion-Format✻ Task	0.299	0.299	0.109	0.362	0.301
Cue-Format✻ Criterion-Format✻ Task	0.114	0.114	0.006	0.020	0.297
Block✻ Cue-Format	0.299	0.299	0.002	0.952	0.002
Block✻ Criterion-Format	0.299	0.299	2.306×10^{-4}	0.477	4.837×10^{-4}
Block✻ Task	0.299	0.299	0.996	1.060×10^{-13}	$9.393 \times 10^{+12}$
Block✻ Cue-Format✻ Criterion-Format	0.114	0.114	2.087×10^{-8}	3.199×10^{-7}	0.065
Block✻ Cue-Format✻ Task	0.114	0.114	0.004	0.002	2.113
Block✻ Criterion-Format✻ Task	0.114	0.114	2.901×10^{-7}	5.555×10^{-5}	0.005
Block✻ Cue-Format✻ Criterion-Format✻ Task	0.006	0.006	1.552×10^{-13}	4.952×10^{-12}	0.031

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt.

Appendix G. Exp. 3 Model comparison of factors influencing performance (RMSE)

Table 1G
Training-Phase Model Comparison Mixed Factorial Bayesian ANOVA.

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.200	9.465e − 29	3.786e − 28	1.000	
Block + Format	0.200	0.729	10.758	7.702e + 27	2.236
Block	0.200	0.207	1.046	2.190e + 27	0.725
Block + Format + Block ✻ Format	0.200	0.064	0.272	6.733e + 26	3.478
Format	0.200	3.247e − 28	1.299e − 27	3.430	1.232

Note. All models include subject. The table starts with the Null Mode. All subsequent models are ordered from the one with the strongest support to the one the weakest support.

Table 2G
Training-Phase Analysis of the Effects of Included Factors.

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF $_{incl}$
Block	0.400	0.400	0.936	4.193e − 28	2.233e + 27
Format	0.400	0.400	0.729	0.207	3.516
Block ✻ Format	0.200	0.200	0.064	0.729	0.087

Note. Compares models that contain the effect to equivalent models stripped of the effect. Higher-order interactions are excluded. Analysis suggested by Sebastiaan Mathôt. The inclusion Bayes Factor (BF$_{incl}$) is the primary factor of concern showing the evidence for including a factor in the final model.

References

Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J., & von Helversen, B. (2020). Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(6), 1064–1090.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull, & R. S. Wyer (Eds.), *Content and process specificity in the effects of prior experiences: Vol. III. Advances in social cognition* (pp. 61–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

Björkman, M. (1973). Inference behavior in nonmetric ecologies. In L. Rappoport, & D. A. Summers (Eds.), *Human judgment and social interaction* (pp. 144–168). Ardent Media.

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. (1980). In one word: Not from experience. *Acta Psychologica, 45*(1), 223–241.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87* (2–3), 137–154.

Brehmer, B., Kuylenstierna, J., & Liljergren, J. E. (1974). Effects of function form and cue validity on the subjects' hypotheses in probabilistic inference tasks. *Organizational Behavior and Human Performance, 11*(3), 338–354.

Bröder, A., Gräf, M., & Kieslich, P. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision making, 12*, 491–506.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments (2nd)*. University of California Press.

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright, & P. Ayton (Eds.), *Judgmental forecasting* (pp. 63–82). John Wiley & Sons.

Castellan, N. J., & Edgell, S. E. (1973). An hypothesis generation model for judgment in nonmetric multiple-cue probability learning. *Journal of Mathematical Psychology, 10* (2), 204–222.

Childers, T. L., & Viswanathan, M. (2000). Representation of numerical and verbal product information in consumer memory. *Journal of Business Research, 47*(2), 109–120.

Collsiöö, A., Sundh, J., & Juslin, P. (2023). Unpacking intuitive and analytic memory sampling in multiple-cue judgment. In K. Fiedler, P. Juslin, & J. Denrell (Eds.), *Sampling in judgment and decision making* (pp. 177–204). Cambridge: Cambridge University Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Academic Press.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95.

De Bock, D., Van Dooren, W., Janssens, D., & Verschaffel, L. (2002). Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics, 50*, 311–313.

De Bock, D., Van Dooren, W., Janssens, D., & Verschaffel, L. (2007). *The illusion of linearity: From analysis to improvement* (Vol. 41). Springer Science & Business Media.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968–986.

D'Errico, J. (2022). *fminsearchbnd, fminsearchcon*. Retrieved December 4, 2017 from https://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd-fminsearchcon. MATLAB Central File Exchange.

Dewolf, T., Van Dooren, W., & Verschaffel, L. (2011). Upper elementary school children's understanding and solution of a quantitative problem inside and outside the mathematics class. *Learning and Instruction, 21*(6), 770–780.

Ebersbach, M., Van Dooren, W., Van den Noortgate, W., & Resing, W. C. (2008). Understanding linear and exponential growth: Searching for the roots in 6-to 9-yearolds. *Cognitive Development, 23*(2), 237–257.

Edgell, S. E. (1983). Delayed exposure to configural information in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Performance, 32*(1), 55–65.

Edgell, S. E., & Castellan, N. J. (1973). Configural effect in multiple-cue probability learning. *Journal of Experimental Psychology, 100*(2), 310–314.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

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

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential brunswik: Beginnings, explications, applications*. Oxford University Press.

von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology, 46*(1), 220–229.

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(4), 867–889.

Hoffmann, J., von Helversen, B., & Rieskamp, J. (2019). Testing learning mechanisms of rule-based judgment. *Decision, 6*(4), 305–344.

Hoffmann, J. A., von Helversen, B., & 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.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(8), 1193–1217.

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review, 28*(5), 1495–1513. https://doi.org/10.3758/s13423-020-01861-1

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., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*(4), 856–874. https://doi.org/10.1037/a0016979

Juslin, P., Nilsson, H., Winman, A., & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition, 120*(2), 248–267.

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.

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

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404–426.

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.

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

Liu, D., Juanchich, M., Sirota, M., & Orbell, S. (2020a). Differences between decisions made using verbal or numerical quantifiers. *Thinking & Reasoning, 0*(0), 1–28.

Liu, D., Juanchich, M., Sirota, M., & Orbell, S. (2020b). The intuitive use of contextual information in decisions made with verbal and numerical quantifiers. *Quarterly Journal of Experimental Psychology (2006), 73*(4), 481–494.

Maciejovsky, B., & Budescu, D. V. (2013). Verbal and numerical consumer recommendations: Switching between recommendation formats leads to preference inconsistencies. *Journal of Experimental Psychology. Applied, 19*(2), 143–157.

Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *The American Journal of Psychology, 93*(3), 429–443.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104–114.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos, & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). New York, NY: Cambridge University Press.

Olsson, A. C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1371.

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

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Scammon, D. (1977). "Information load" and consumers. *Journal of Consumer Research, 4*, 148–155.

Schkade, D. A., & Kleinmuntz, D. N. (1994). Information displays and choice processes: Differential effects of organization, form, and sequence. *Organizational Behavior and Human Decision Processes, 57*(3), 319–337.

Stone, D. N., & Schkade, D. A. (1991). Numeric and linguistic information representation in multiattribute choice. *Organizational Behavior and Human Decision Processes, 49*(1), 42–59.

Sundh, J., Collsiöö, A., Millroth, P., & Juslin, P. (2021). Precise/not precise (PNP): A Brunswikian model that uses judgment error distributions to identify cognitive processes. *Psychonomic Bulletin & Review, 28*, 351–373.

Trippas, D., & Pachur, T. (2019). Nothing compares: Unraveling learning task effects in judgment and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(12), 2239–2266.

Van Dooren, W., De Bock, D., Depaepe, F., Janssens, D., & Verschaffel, L. (2003). The illusion of linearity: Expanding the evidence towards probabilistic reasoning. *Educational Studies in Mathematics, 53*, 113–138.

Van Dooren, W., De Bock, D., Janssens, D., & Verschaffel, L. (2008). The linear imperative: An inventory and conceptual analysis of students' overuse of linearity. *Journal for Research in Mathematics Education, 39*(3), 311–342.

Verschaffel, L., De Corte, E., & Lasure, S. (1994). Realistic considerations in mathematical modelling of school arithmetic word problems. *Learning and Instruction, 4*, 273–294.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*(2), 135–138.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364.