# Competitive Retrieval Strategy Causes Multimodal Response Distributions in Multiple-Cue Judgments

Rebecca Albrecht[a*], Janina A. Hoffmann[b], Timothy J. Pleskac[c,d], Jörg Rieskamp[a], Bettina von Helversen[e,f]

[a] University of Basel, Missionsstrasse 62A, 4055 Basel Switzerland

[b] University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany

[c] University of Kansas, Lawrence, Kansas, USA

[d] Max Planck Institute for Human Development, Berlin, Germany

[e] University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland

[f] University of Bremen, Hochschulring 18 Cognium Building, 28359 Bremen, Germany

## Author Note

[*] Corresponding author: rebecca.albrecht@unibas.ch

The data and the models are available at:
`https://osf.io/96m8g/?view_only=9371ff38c79c4fb08a539d3156266b5c`

We thank Regina Weilbächer and Florian Seitz for their support in collecting the data and Anita Todd for proof-reading the manuscript.

This research has been supported by the Swiss National Science Foundation (SNSF) grant no. 146169 to the fourth and the last author and grant no. 157432 to the last author.

Author contribution R.A., J.H., T.P., J.R., B.vH. designed the research. R.A, J.H, B.vH. conceptualized the cognitive models and experimental design. R.A. performed the data analysis, implemented the experiments, and programmed the cognitive models. R.A, J.H., B.vH. wrote the original draft. R.A., J.H., B.vh, T.P., J.R. edited and reviewed the manuscript.

## Abstract

Research on quantitative judgments from multiple cues suggests that judgments are simultaneously influenced by previously abstracted knowledge about cue–criterion relations and memories of past instances (or exemplars). Yet extant judgment theories leave two questions unanswered: (a) How are past exemplars and abstracted cue knowledge combined to form a judgment? (b) Are all past exemplars retrieved from memory to form the judgment (integrative retrieval) or is the judgment based on one exemplar (competitive retrieval)? To address these questions we propose and test a new model, CX-COM (combining **C**ue abstraction with e**X**emplar memory assuming **COM**petitive memory retrieval). In a first step, CX-COM recalls only a single exemplar from memory. In a second step, the initially retrieved judgment is adjusted based on abstracted cue knowledge. Qualitatively, we show that CX-COM naturally captures judgment patterns that have been previously attributed to multiple strategies. Next, we tested CX-COM quantitatively in two experiments and found that it accounts well for people's judgment behavior. In the second experiment we additionally tested two qualitative predictions of CX-COM: The existence of multimodal response distributions within participants and systematic variability in judgments depending on the distance between similar exemplars in memory. The empirical results confirm CX-COM's assumptions. In sum, the evidence suggests that CX-COM is a viable new model for quantitative judgments and shows the importance of considering judgment variability in addition to average responses in judgment research.


*Keywords:* Quantitative judgment, multiple cues, exemplar retrieval, cue abstraction, retrieval theory, mixture models

# Competitive Retrieval Strategy Causes Multimodal Response Distributions in Multiple-Cue Judgments

## Introduction

Evaluating situations and judging the value of objects is a widespread cognitive task carried out every day in people's professional and private lives. From a judge passing sentence on a convict to a financial analyst evaluating the risk and value of a bond or stock, people's ability to estimate numerical criteria in many different domains is of high importance. When making judgments people use the information of different features or attributes (cues) describing an object or situation. A judge determining the length of a sentence for a robbery conviction, for example, might consider the extent of the damages in the case. To do so, the judge might retrieve details of past cases from memory and compare them to the facts of the current case. Such a judgment strategy is usually described by exemplar models (Nosofsky, 2014). These models assume that people's judgments and decisions are based on the similarity between the object under consideration and exemplars stored in memory (Hoffmann, von Helversen, & Rieskamp, 2014; Juslin, Jones, Olsson, & Winman, 2003; Juslin, Olsson, & Olsson, 2003; Nosofsky, 1984, 1986, 1997). Exemplar models have been successfully used to explain a variety of phenomena across different domains ranging from memory recall (e.g. Brown, Neath, & Chater, 2007; Hintzman, 1984) to categorizations and classifications (Medin & Schaffer, 1978; Nosofsky, 1984) to decision making (Juslin & Persson, 2002; Pachur & Olsson, 2012; Platzer & Bröder, 2012). They have also been extended to account for judgments from multiple cues (Hoffmann et al., 2014; Hoffmann, von Helversen, & Rieskamp, 2016; Juslin, Jones, et al., 2003; Juslin, Karlsson, & Olsson, 2008; von Helversen & Rieskamp, 2009).

Despite their success in cognitive psychology, approaches for quantitative judgments that are purely based on exemplar processing fail to address two problems: The first problem is that people learn to explicitly represent how cues relate to a criterion and use this knowledge to make predictions for new objects (Brehmer, 1994; Cooksey, 1996; Juslin, Jones, et al., 2003). Following such a cue-abstraction process, the judge, returning to our earlier example, would pass a prison sentence in a robbery case by weighing the importance of the aggravating and mitigating factors (e.g. the damages caused and whether the robber showed remorse) and then combining the weighted factors to form a single sentence. Cue-abstraction processes are hard to reconcile with exemplar-based strategies. As a consequence, current research in judgment assumes that people rely on both processes but switch between them depending on the structure of the task and their own cognitive abilities (Herzog & von Helversen, 2018; Hoffmann et al., 2014, 2016; Juslin, Jones, et al., 2003; Juslin et al., 2008; Juslin, Olsson, & Olsson, 2003; Pachur & Olsson, 2012; von Helversen & Rieskamp, 2008, 2009). Yet, empirical evidence does not unequivocally favor the view that cue abstraction proceeds independently

of exemplar retrieval. For instance, the similarity between the to-be-judged event and past instances influences people's judgments even when they are relying on rules and abstracted knowledge (Brooks & Hannah, 2006; Hahn, Prat-Sala, Pothos, & Brumby, 2010; von Helversen, Herzog, & Rieskamp, 2014). Still, it is an open question how people integrate the two types of processes, that is, cue-abstraction and exemplar-based processes, to form a judgment, which we address in the present work.

The second problem is that although exemplar models are deeply rooted in traditional models of memory, how exemplar models instantiate retrieval from memory diverges from the retrieval processes considered in contemporary memory models. Specifically, exemplar models in quantitative judgment assume an integrative retrieval mechanism where all previously encountered exemplars are activated in parallel and integrated into one composite value (Hoffmann et al., 2016; Pachur & Olsson, 2012). In contrast to this view, many contemporary memory models assume a competitive retrieval process where previously encountered exemplars compete for retrieval and only one exemplar is recalled (Anderson, 1983; Logan, 1988). The degree to which a competitive retrieval mechanism better captures how people retrieve past exemplars during the judgment process has not been investigated.

The goal of the present research was to propose and test an exemplar-based model for quantitative judgments that addresses these two limitations. CX-COM (combining **C**ue abstraction with e**X**emplar memory assuming **COM**petitive memory retrieval) proposes that people engage in a two-step judgment process: In the first step, people probabilistically retrieve one past exemplar from memory, and in the second step, they adjust the criterion value of the recalled exemplar based on knowledge about the cue–criterion relation. In the present work we first present evidence for different retrieval and knowledge integration mechanisms and their implications for the judgment process. Next, we formally derive the predictions of CX-COM from established versions of exemplar and cue-abstraction models and review how the new model can account for behavioral patterns frequently observed in multiple-cue judgment. We then present two experiments that (a) quantitatively test CX-COM against competing models (Experiment 1) and (b) test the qualitative prediction that previously learned exemplars compete for retrieval (Experiment 2). In the General Discussion we compare CX-COM to cognitive models proposed for categorization and function learning and discuss limitations and potential future work.

## Combining exemplar and cue-abstraction processes

Exemplar models propose that people represent learned instances by storing them in memory. Alternatively, cue-abstraction accounts propose that people conceptualize learned knowledge on a more abstract level as a set of rules or cue–criterion relations (Hahn & Chater, 1998; Macrae et al., 1998). The general question of how knowledge is represented and used has challenged judgment research over the past decade (Hoffmann et al., 2014, 2016; Juslin, Jones, et al., 2003; Juslin et al., 2008; Karlsson,

Juslin, & Olsson, 2007; von Helversen et al., 2014; von Helversen & Rieskamp, 2008, 2009). By analyzing judgment behavior in different domains, it has been shown that the judgment process varies with the structure of the judgment task and the cognitive abilities of the decision maker. For instance, participants are better described by cue-abstraction models if the judgment criterion is a linear, additive function of the cues, whereas nonlinear relationships are better captured with exemplar models (Hoffmann et al., 2016; Juslin et al., 2008). In this vein, current research in judgment portrays the judgment process as a selection from two types of judgment processes best described by rule-based cue-abstraction models or similarity-based exemplar models (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012). This research implicitly suggests that people first select one process suited to the task at hand and then use only the output of this process to make judgments in the task.

Empirical evidence, however, suggests that exemplar retrieval and abstracted cue knowledge likely interact during categorization and judgment. Unintentionally activated exemplars can interfere with task performance if they do not match the demands of the current situation (e.g. Macrae et al., 1998). Specific exemplars can influence judgments that are otherwise based on abstracted cue knowledge (von Helversen et al., 2014) and activate different rules depending on the context in which past exemplars were learned (e.g. Yang & Lewandowsky, 2004). Consequently, categorization research tends to favor mixture or hybrid models that assume people's representations contain both generalized beliefs in the form of abstracted (cue) knowledge and specific instances or exemplars. In these models both types of representations influence decisions, although the relative importance may differ depending on the task and learning history (e.g. Anderson & Betz, 2001; Ashby, Alfonso-Reese, et al., 1998; Erickson & Kruschke, 1998; Herzog & von Helversen, 2018; Nosofsky, Palmeri, & McKinley, 1994; Palmeri, Wong, & Gauthier, 2004; Vanpaemel & Storms, 2008).

Thus, it seems reasonable to assume that people integrate retrieval from memory with abstracted cue knowledge also in multiple-cue judgment. But how do these two processes interact? In categorization research different types of mixture and hybrid models have been proposed. Most prominently in blending models, an exemplar and a cue-abstraction mechanism process information in parallel and the two outputs are combined as a weighted average (e.g. Bröder, Gräf, & Kieslich, 2017; Erickson & Kruschke, 1998). The most recent implementation of a blending model for judgments is RulEx-J that captures the contribution of exemplar- and rule-processing across different task conditions (Bröder et al., 2017).

## Integrative versus competitive retrieval

Any model of memory has to address the key question of how exemplars stored in memory are retrieved, that is activated and recalled. Memory models often share the assumption that the activation of exemplars is based on their similarity to the current stimulus (the probe). They differ,

however, in the way a recalled exemplar is produced upon request (Raaijmakers & Shiffrin, 1992). Two different approaches can be distinguished, an integrative and a competitive retrieval mechanism.

An integrative retrieval mechanism produces a composite of all exemplars (or of a subset) in memory. The most prominent example for an integrative retrieval mechanism is employed in the MINERVA model (Dougherty, Gettys, & Ogden, 1999; Hintzman, 1984). In this model, exemplars are represented as feature lists called memory traces. A probed recall activates all memory traces in parallel and yields a special memory trace: an echo. This echo is the sum of all traces in memory, each weighted by its activation value. Similarly, memory models that assume composite storage, such as TODAM (Lewandowsky, Murdock, et al., 1989), usually also yield a composite as a retrieval product.

If past exemplars compete for retrieval only one exemplar is produced on each retrieval attempt. This competitive retrieval mechanism can be found in a number of established memory models, such as the ACT-R theory (Anderson, 1983), the instance theory of automatization (Logan, 1988, 2002), the SAM model (Raaijmakers & Shiffrin, 1980), and some random walk theories (Nosofsky, 1997; Ratcliff, 1978). A competitive retrieval mechanism assumes that exemplars in memory are stored and accessed separately. Although some of these theories specify how subsequently retrieved exemplars can be combined to form a task response, they assume that each retrieval request yields only one exemplar.

The retrieval mechanism employed (integrative vs. competitive) implies different response processes (Juslin & Persson, 2002; Palmeri, 1997). Exemplar models in categorization or judgment mostly postulate an integrative retrieval mechanism (e.g. Hoffmann et al., 2014; Juslin, Olsson, & Olsson, 2003; Medin & Schaffer, 1978; Nosofsky, 1984; Pachur & Olsson, 2012). Once a probe is presented, all exemplars stored in memory are activated and a judgment or category response is formed as a weighted average over all memory items (Hoffmann et al., 2014; Juslin et al., 2008; Juslin, Olsson, & Olsson, 2003; Medin & Schaffer, 1978; Nosofsky, 1984). One potential reason why exemplar models seldom consider competitive retrieval is that often (at least in the domains that have been considered) the two mechanisms predict the same responses. In categorization tasks, for instance, classic exemplar models predict the probability of a new item belonging to a category by using the sum of similarities it holds with exemplars in that category. A competitive retrieval mechanism predicts that in each trial, an exemplar is recalled from memory and used as a basis for the category decision. However, the sum of the recall probabilities of individual exemplars belonging to a category is the same as the probability of assigning an item to a category if integrative retrieval is assumed. Thus, the two retrieval mechanisms cannot be distinguished in this type of task. In quantitative judgment tasks the type of retrieval can be important because integrative and competitive retrieval mechanisms make qualitatively distinct predictions on the judgment level.

Within the domain of judgments, integrative and competitive retrieval can be distinguished by the predicted trial-by-trial variability across items and (sometimes) different distribution shapes. An

integrative retrieval mechanism predicts that all exemplars in memory are combined into a single response value. Within-participant trial-by-trial variability is typically assumed to be normally distributed (e.g. Pachur & Olsson, 2012; Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009), resulting in a model-predicted *unimodal response distribution* centered around the predicted response value. With competitive retrieval each exemplar can, in principle, be recalled, although its chances in a given context might be very low. As a result, a competitive retrieval mechanism predicts *multimodal response distributions* and systematic changes in across-item variability. A detailed example will be discussed in the next section.

## CX-COM: A hybrid model for quantitative judgment with a competitive retrieval mechanism

The development of CX-COM was motivated by two currently unresolved questions in judgment research: (a) When forming a judgment based on exemplars retrieved from memory, does the retrieval request yield one exemplar or an integrative composite of all exemplars? (b) Does combining cue-abstraction processes with exemplar retrieval outperform the predictions of a pure exemplar-based or cue-abstraction process?

CX-COM addresses these two questions by proposing a two-step judgment process: In the first step, previously encountered exemplars compete for retrieval and only the winning exemplar along with its criterion value is recalled from memory. Second, a cue-based adjustment process uses the recalled exemplar as a reference point and adjusts the criterion value depending on the generalized beliefs about cue-criterion relations. Accordingly, CX-COM spells out how people may combine competitive exemplar-retrieval and cue-abstraction processes, allowing one to test these assumptions against single-process models as well as competing mixture models.

In this section, we first introduce established models of human judgment, that is, classical exemplar and cue-abstraction models. Next, we explain CX-COM and its components in relation to the established models. We also introduce the blending model RulEx-J (Bröder et al., 2017) as an additional competitor. A running example at the end of each subsection highlights differences and similarities in the models' predictions. As a last step we review important findings from the judgment literature and explain how CX-COM accounts for them.

### Exemplar models

In exemplar models (Nosofsky, 2014), a probe $p$ that has to be judged or categorized serves as a retrieval cue, activating previously encountered exemplars in memory. A response $\hat{j}_p^{Exemplar}$ is an average of all judgment values $j_e$ associated with exemplars $e$ in the set of all exemplars $M$ weighted by

their relative, subjective similarity to $p$,

$$\hat{j}_p^{\text{Exemplar}} = \frac{\sum_{e \in M} sim(e,p) \cdot j_e}{\sum_{e \in M} sim(e,p)}. \tag{1}$$

The more similar an exemplar in memory is to the probe, the higher its impact on the response value. The similarity between exemplars $e$ and probe $p$ depends exponentially on their distance in psychological space,

$$sim(e,p) = e^{-c \cdot dist(e,p)}. \tag{2}$$

Parameter $c$ is the sensitivity parameter and manipulates how much impact the psychological distance between a probe and an exemplar has on the subjective perception of similarity. Lower values of $c$ imply that two items with a high distance in psychological space are still perceived as similar.

The distance in psychological space is usually described using the family of Minkowski distance metrics. For an exemplar $e$ with $n$ cue dimensions and cue values $c_1^e, \ldots, c_n^e$, a probe $p$ with cue values $c_1^p, \ldots, c_n^p$, and attention weights $w_1, \ldots, w_n$ this would be

$$dist(e,p) = \sqrt[r]{\sum_{i=1}^{n} w_i \cdot |c_i^e - c_i^p|^r}. \tag{3}$$

Attention weights are assumed to vary between 0 and 1 and are constrained to sum to 1. The parameter $r$ captures how visually distinguishable cue dimensions are in a given task. The so-called city-block distance is defined by $r = 1$ and is used when the dimensions are very distinct (Garner, 2014; Shepard, 1964). The Euclidean distance is represented by $r = 2$ and is used when the dimensions overlap.

The response value $\hat{j}_p^{\text{Exemplar}}$ is a composite of the exemplars in memory and is predicted each time probe $p$ is presented. Usually, a normally distributed error is associated with a response. Thus, the predicted distribution of criterion values, response distribution $R_p$, coincides with the assumed error distribution and is for one item $p$ and some variance $\sigma^2$:

$$R_p^{\text{Exemplar}} \sim \mathcal{N}(\hat{j}_p^{\text{Exemplar}}, \sigma^2). \tag{4}$$

Competitive exemplar models make the same assumptions about the psychological distance (Equation 3) and similarity between the exemplars in memory and a probe (Equation 2). However, the relative similarity now determines the probability to recall exemplar $e$ given probe $p$ so that

$$pr(e|p) = \frac{sim(e,p)}{\sum_{e \in M} sim(e,p)}. \tag{5}$$

In each trial only one exemplar $e$ is recalled from memory and the associated criterion value $j_e$ is given as a response, i.e. $j_p^{\text{Exemplar-competitive}} = j_e$. In another trial, an exemplar with another criterion value may be retrieved probabilistically so that this competitive retrieval elicits a multimodal response distribution. Assuming also a normally distributed error associated with a response in every trial, the

response distribution $R_p^{\text{Exemplar-competitive}}$ is a mixture of normal distributions with modes given by the criterion values $j_e$ stored in memory:

$$R_p^{\text{Exemplar-competitive}} \sim \sum_{e \in M} pr(e|p) \cdot \mathcal{N}(j_p^{\text{Exemplar-competitive}}, \sigma^2). \quad (6)$$

Thus, the predicted response distribution is quite different compared to the one predicted by the integrative exemplar model.

***Example.*** To illustrate how an exemplar model with an integrative retrieval mechanism (Equation 4) and an exemplar model with a competitive retrieval mechanism (Equation 6) make different predictions in quantitative judgments, consider a judge passing sentence on a bank robber (see also Table 1). In an attempt to rob a bank, a robber (Defendant 1) caused low property damage and low harm to people. The judge can relate these circumstances to two earlier cases, one with low property damage (but high harm to people) and a prison sentence of 8 years, and another with low harm to people (but high property damage) and a sentence of 4 years. Assuming equal importance of both aspects of the case ($w_{harmtopeople} = w_{propertydamage}$), the sentence would be 6 years (because both old cases are equally similar to the new case). Assuming normally distributed deviations from the recalled criterion value in the model, repeated sentencing would result in a unimodal distribution of judgments centered around 6 years (see Figure 1, *Exemplar (integrative)*). In contrast, an exemplar model with a competitive retrieval mechanism (without additional assumptions on the response process) predicts sometimes a prison sentence of 4 years and sometimes a sentence of 8 years, depending on which of the two earlier cases the judge recalls. Assuming also that deviations from a recalled criterion value happen by chance, the judge would draw the prison sentence from a bimodal distribution with modes at criterion values 4 and 8 (see Figure 1, *Exemplar (competitive)*). Both integrative and competitive exemplar models would not predict a sentence below 4 years for Defendant 1 despite the fact that he caused less harm or less damage than the remembered convicts.

Consider in comparison Defendant 2 who caused high property damage and a high harm to people. Assuming equal dimension weights, both an integrative and a competitive exemplar model would predict the same sentence as for Defendant 1. The integrative exemplar models yields a unimodal sentence around 6 years; the competitive exemplar model a multimodal sentence, retrieving 4 and 8 years.

However, note that multimodal response distributions will not always occur with a competitive retrieval mechanism, but depend on the similarity structure of the training set given a probe, in particular the attention to specific dimensions. For example, assume again Defendant 2, who caused *high* harm to people and high property damage. However, the judge does not equally weight the two dimensions but only considers the harm to other people and neglects property damage (e.g. w1=2, w2=100). In this case, convict 2 will be perceived as much more similar to Defendant 2 than Convict 1

Table 1

*Bank robber example*

| | Memory | | Novel | |
|---|---|---|---|---|
| | Convict 1 | Convict 2 | Defendant 1 | Defendant 2 |
| Property damage | *High* | Low | Low | High |
| Harm to people | Low | *High* | Low | High |
| Sentence | 4 | 8 | TBD | TBD |

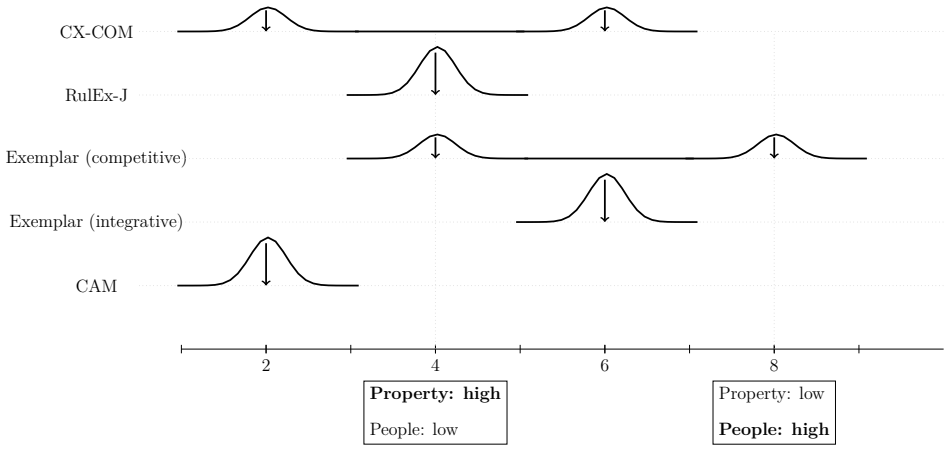


*Figure 1*. Example (see Table 1) showing the response distributions predicted by the cue-abstraction model (CAM), the exemplar model, RulEx-J, and the cue abstraction model with exemplar memory assuming competitive memory retrieval (CX-COM) for Defendant 1 (low property damage and low harm to people).

and will be retrieved with a much higher likelihood. As a result, the judge may always pass a sentence of 8 years and, consequently, a unimodal response distribution would emerge.

## Cue-abstraction models

Cue abstraction models propose that people extract and explicitly represent their beliefs about the importance of cues as a set of weights. Cue-abstraction models (CAMs) for judgments are often implemented as main effects linear regression models (Juslin et al., 2008), because this implementation has been shown to fit judgment data especially well in a variety of domains (for reviews see Karelaia & Hogarth, 2008; Kaufmann, Reips, & Wittmann, 2013). To make a judgment, cue values $c_1 \ldots c_n$ of a

probe $p$ are weighted by their relative importance $b_i$ and summed up so that

$$\hat{j}_p^{\text{CAM}} = k + \sum_{i=1}^{n} (b_i \cdot c_i). \tag{7}$$

Parameter $k$ is an intercept, e.g. the baseline judgment in case of cue values of zero. Similar to exemplar models, the CAM's prediction remains the same over repeated presentations of probe $p$ and a normally distributed error is assumed resulting in the response distribution $R_p^{\text{CAM}}$ centered around the response value $j_p^{\text{CAM}}$ (similar to Equation 4).

***Example.*** Figure 1 shows again an example of the unimodal response distribution predicted by the CAM in the bank robber's case (Table 1, Defendant 1). Assuming a minimum sentence of two years for a robbery attempt with low property damage and low harm to people, the judge could weigh high property damage as an additional 2 years of prison and high harm to people as an additional 6 years[1]. This would result in a prison sentence for the novel Defendant 1 of 2 years. For Defendant 2 (high property damage and high harm to people) the CAM with the same assumptions would thus predict a unimodal response distribution with a mode at 10 years.

## Combining competitive exemplar retrieval with cue-abstraction

With CX-COM we propose a two step process: First, an exemplar is recalled from memory. Second, the associated criterion value is adjusted based on the beliefs about the cue–criterion relationship. Exemplar retrieval follows the same principles as in the exemplar model with a competitive retrieval mechanism. Following the presentation of a probe $p$, all exemplars $e$ in exemplar memory $M$ are activated based on their relative, subjective similarity with similarity and psychological distance calculated as described in Equations 2 and 3. The relative similarity determines the probability to recall exemplar $e$ and is described in Equation 5.

In each trial one exemplar is recalled from memory. This exemplar is used as a reference point for a cue-abstraction process. Specifically, a cue-based adjustment mechanism adjusts the criterion

[1]Note that we use different dimension weights for the exemplar model and the CAM to illustrate how the mechanisms assumed by the models can lead to differential predictions. If we assume the same dimension weights in the exemplar model as in the CAM, i.e. $w_{harmtopeople} = 6$ and $w_{propertydamage} = 2$, the exemplar models become more difficult to distinguish. The integrative exemplar model predicts a unimodal distribution centered around a prison sentence of 4.1 years for Defendant 1 and centered around 7.9 years for Defendant 2. The exemplar model with competitive retrieval predicts a bimodal distribution with modes at the two sentences of the previous cases 4 and 8. However, for Defendant 1 the judge has a 98% chance to recall the case of Convict 1 and for Defendant 2 he has a 98% to recall the case of Convict 2. For CX-COM we discuss how its predictions depend on its parameters in the section "Predicting multiple-cue judgments with CX-COM".

value $j_e$ of a recalled exemplar $e$ to obtain a response. The magnitude of the change depends on the differences in cue values between the probe $p$ and the exemplar $e$ on each cue dimension and the relative importance given to the cue dimension, represented by a cue dimension weight $b_i$:

$$\hat{j}_p^{\text{CX-COM}} = j_e + \left( \sum_{i=1}^{n} b_i \cdot (c_i^e - c_i^p) \right) \cdot \alpha. \tag{8}$$

Parameter $\alpha$ is a scaling parameter that reflects how much the observed difference in cue values influences the judgment, and $c_i^p$ and $c_i^e$ denote the cue values of the probe and the recalled exemplar.

The competitive retrieval in CX-COM elicits a response distribution quite different from the integrative exemplar model and the cue abstraction model. Assuming a normally distributed error associated with a response in every trial, the response distribution $R_p^{\text{CX-COM}}$ is a mixture of normal distributions with modes close to criterion values $j_e$ stored in memory and weighted by the similarity of the associated exemplars $e$ to the probe $p$:

$$R_p^{\text{CX-COM}} \sim \sum_{e \in M} pr(e|p) \cdot \mathcal{N}(\hat{j}_p^{\text{CX-COM}}, \sigma^2). \tag{9}$$

***Example.*** Figure 1 also shows an example of the multimodal response distribution predicted by CX-COM in the bank robber's case (Table 1, Defendant 1). Assuming equal importance of both aspects of the case, the similarity between Defendants 1's case and the two older cases is the same and so is their chance to be recalled, similar to the predictions of the exemplar model with competitive retrieval. In CX-COM, however, the modes of the multinomial response distribution are shifted depending upon the difference in cue values between the current case and the recalled case. Making the simplified assumption that a difference between high and low damage/harm is 2 years of prison, the modes would be adjusted downwards by 2 years each from 4 years to 2 years and from 8 years to 6 years, respectively. For Defendant 2, however, the predicted modes are adjusted upwards from 4 to 6 and from 8 to 10, respectively. Thus, because the cue abstraction component is sensitive to the direction of the adjustment, CX-COM also predicts different response distributions from an exemplar model with only competitive retrieval.

## Blending Models

Besides, pure exemplar and cue abstraction models, blending models such as ATRIUM for categorization (Erickson & Kruschke, 1998) and the measurement model RulEx-J for judgments (Bröder et al., 2017) have been proposed. These models assume an independent processing of exemplar and cue-abstraction models with the overall response being a weighted average (or blend) of the predictions of the single responses. As it is the most recent blending model for judgments, we included RulEx-J in the model test.

Given the judgments for probe $p$ predicted by the exemplar model, $\hat{j}_p^{\text{Exemplar}}$, and the CAM, $\hat{j}_p^{\text{CAM}}$, the response of RulEx-J is

$$\hat{j}_p^{\text{RulEx-J}} = \beta \cdot \hat{j}_p^{\text{CAM}} + (1 - \beta) \cdot \hat{j}_p^{\text{Exemplar}} \tag{10}$$

with the parameter $\beta$ weighting the relative contribution of each model's response. The response distribution $R_p^{\text{RulEx-J}}$ is unimodal as in the exemplar model and the CAM, but the mode lies between the predictions of these models:

$$R_p^{\text{RulEx-J}} \sim \mathcal{N}(\hat{j}_p^{\text{RulEx-J}}, \sigma^2). \tag{11}$$

Although they are both mixture models, the blending model differs from CX-COM in two important aspects: (1) the blending model assumes a parallel processing of an exemplar and a cue-abstraction component while CX-COM assumes that cue abstraction acts upon the retrieved exemplar and (2) the blending model's exemplar component assumes integrative retrieval while CX-COM assumes competitive retrieval.

***Example.*** According to RulEx-J, the judge sentences Defendant 1 (Table 1) to a mixture of the predictions of the Exemplar model with integrative retrieval and the CAM. Making the same assumptions for the two models as in the respective examples and additionally assuming that the judge gives equal weights to both models, the sentence would be be 4 years, exactly the middle between the exemplar model's prediction (6 years) and the CAM's prediction (2 years).

## Predicting multiple-cue judgments with CX-COM

In the past decade, research on multiple cue judgments has proposed that judgment strategies may elicit distinct behavioral judgment patterns (Bröder et al., 2017; Hoffmann et al., 2014, 2016; Juslin et al., 2008; Mata, von Helversen, Karlsson, & Cüpper, 2012; Pachur & Olsson, 2012; von Helversen, Mata, & Olsson, 2010; von Helversen & Rieskamp, 2009). For instance, Juslin et al. (2008) showed that in a linear judgment task people showed extrapolation, that is they judged probes with lower/higher cue values than the training exemplars as having lower/higher criterion values than the training exemplars. This judgment pattern matches the predictions of the CAM (Figure 2a), but disagrees with the predictions of an exemplar model in that task (Figure 2b). In contrast, in a multiplicative environment participants do not seem to extrapolate beyond the range of encountered training values and thus participants' responses match the predictions of the exemplar model (Figure 2e), but disagree with the CAM's predictions (Figure 2d). In general, these differences in judgment patterns have been taken as evidence that people shift between exemplar memory and cue abstraction processes (Bröder et al., 2017; Hoffmann et al., 2014, 2016; Juslin et al., 2008; von Helversen et al., 2010).

CX-COM does not assume a shift between judgment processes, but proposes that a cue adjustment process acts on the retrieved exemplars. The *alpha* parameter governs the extent to which

the retrieved criterion value is adjusted based on cue knowledge. In the following we show that CX-COM can capture the same behavioral judgment patterns that have been reported in the literature without assuming a change in judgment processes and analyze how different parameter settings influence CX-COM's predictions.

In a first step, we generated CX-COM's predictions for the linear and the multiplicative judgment task reported by Juslin et al. (2008), using the reported parameters for the CAM as dimension weights and an additive similarity function with equivalent parameters as attention weights for the exemplar model. Figure 2c and 2f illustrates that CX-COM predicts a similar change in judgments depending on the task structure as predicted by a strategy shift. In the linear environment its predictions resemble the predictions of the CAM, whereas its predictions lie between the exemplar model and the CAM in the multiplicative environment. Thus CX-COM reflects participants' responses in both environments. Notably, CX-COM accounts for these behavioral patterns without adjusting any parameter values across environments but the changes result from differences in the structure of the environment. One reason is that in a linear task with correct weights the adjustment process by CX-COM leads to the same judgment independent of which exemplar was retrieved.[2] In contrast, in a multiplicative task predictions will differ depending on the retrieved exemplar leading to exemplar effects and reducing extrapolation on the average level.

But can CX-COM also explain strategy shifts within the same environment due to within-task manipulations such as instructions or individual preferences? To understand whether the free parameters in CX-COM allow it to capture exemplar-based and cue-based judgment patterns within a task, we analyzed within Juslin et al.'s linear environment (Juslin et al., 2008) how changes in the parameter values, specifically changes in the adjustment parameter $\alpha$ and in the dimension weights, influence CX-COM's predictions. Assuming correct dimension weights with parameter $\alpha = 1$ (Figure 3e), CX-COM produces the exact same predictions as the CAM (Figure 3a), and CX-COM with $\alpha = 0$ (Figure 3f) produces the exact same predictions as the exemplar model (Figure 3b). Assuming incorrect weights, for instance uniform weights equal to 1, the predictions of CX-COM for $alpha = 0$ (Figure 3h) are still exactly the same as for the exemplar model (Figure 3d). However, with $\alpha = 1$ the predictions differ between the CAM (Figure 3c) and CX-COM (Figure 3g), with CX-COM showing a judgment pattern that deviates from the CAM with incorrect weights, but is still linear. Accordingly, on average, if $alpha = 0$ CX-COM reduces to an exemplar model with the same dimension weights.

[2]For example, assume training items $(c_1 = 1, c_2 = 1)$ with criterion 4 and $(c_1 = 2, c_2 = 2)$ with criterion 8. Assume further that a participant abstracts the (correct) cue weights w1 = w2 = 2 from these items. For a test item $(c_1 = 1, c_2 = 2)$ the response would be 6, independently of whether the first training item is recalled and the associated criterion value is increased or if the second training item is recalled and its criterion value is decreased.

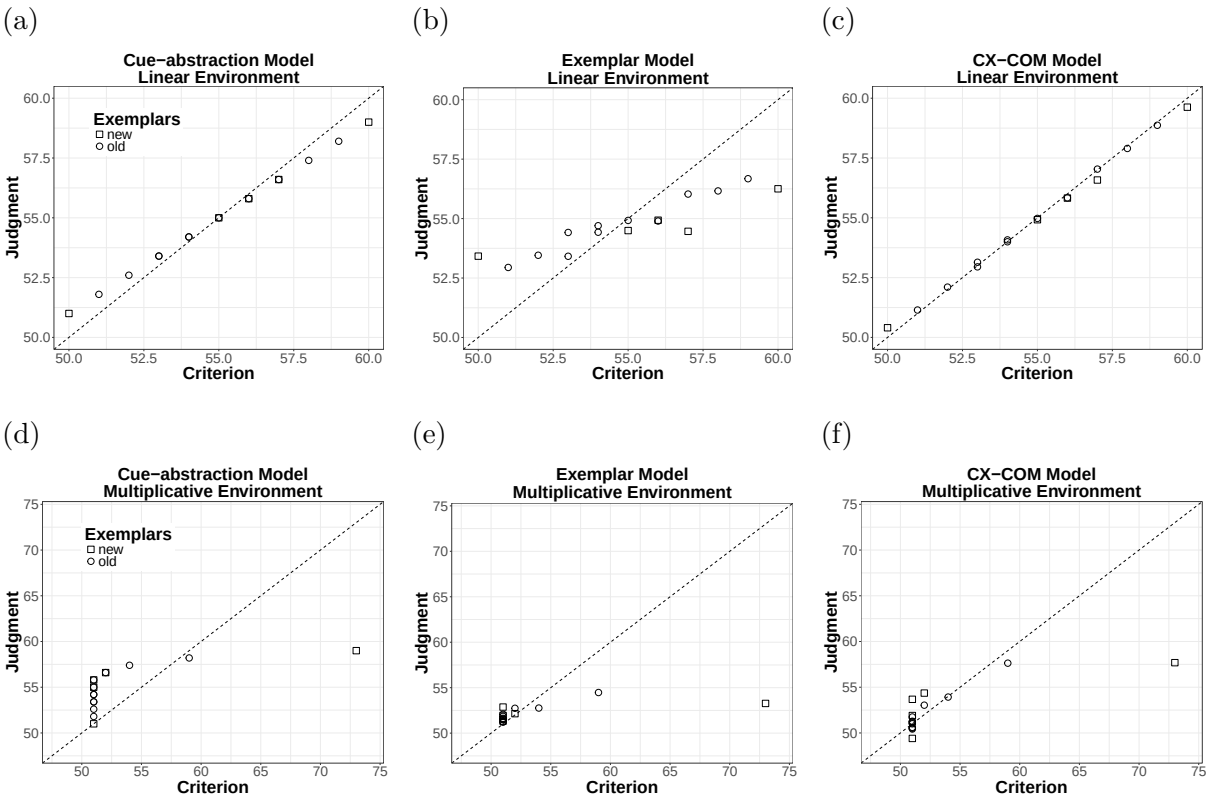


*Figure 2*. Shows predictions of a cue-abstraction model (panels a and d), an exemplar model (panels b and e) and CX-COM (panels c and f) in a linear environment (panels a-c) and a multiplicative environment (panels d-f). Items, models, and parameters are the same as described in Juslin et al. (2008). For CX-COM we used the same weights as for the linear model ($w_1 = 3.2$, $w_2 = 2.4$, $w_3 = 1.6$, $w_4 = 0.8$) and $\alpha = 1$.

With increasing *alpha*[3], i.e. more cue adjustment, the predictions become more linear resembling a rule-based process. This suggests that *alpha* in CX-COM reflects the extent to which participants' judgments are influenced by cue-abstraction processes, similar to the interpretation of alpha in other mixture models like Rulex-J. However, the CAMs and CX-COMs predictions are only identical when the correct cue weights are assumed.

These simulations indicate that CX-COM can also reflect different levels of exemplar and cue-abstraction processes induced by manipulations within a task environment. For instance, previous research has argued that changing only one cue between subsequent exemplars facilitates cue abstraction processes (Juslin et al., 2008). Within CX-COM, this could be reflected by a stronger reliance on adjustment processes resulting in a lower $\alpha$ parameter for confounded than for ordered sequences.

[3]please note that *alpha* as well as the dimension weights depend on the judgment scale and thus can not be easily compared between tasks

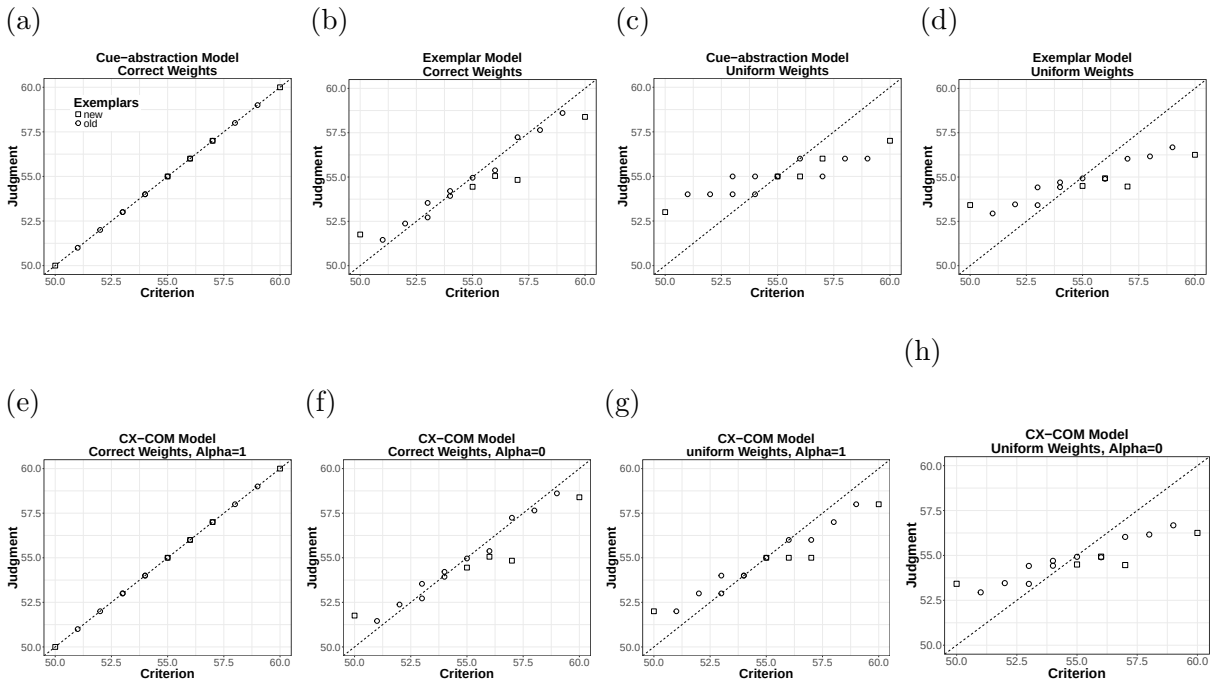


*Figure 3*. Predictions of the CAM (a,c), the exemplar model (b,d) and CX-COM (e-h) assuming correct weights ($w_1 = 1$, $w_2 = 2$, $w_3 = 3$, $w_4 = 4$; panels a,b,e,d) and incorrect uniform weights ($w_1 = 1$, $w_2 = 1$, $w_3 = 1$, $w_4 = 1$; panels c,d,g,h). Sensitivity Parameter c in all models including an exemplar component is set to the sum of the weights and attention weights are set to the sum of weights divided by c. Panels e and g show CX-COM's predictions assuming a strong influence of the cue-abstraction process (i.e. $\alpha = 1$) and panels f and h shows CX-COM's predictions assuming a pure exemplar process (i.e. $\alpha = 0$). When assuming correct weights CX-COM perfectly mimics the exemplar model's and the CAM's predictions depending on the value of parameter $\alpha$ (compare panels a and e, and panels b and f). Assuming uniform weights CX-COM with a $\alpha$ of 0 still matches the exemplar model's predictions (compare panels d and h), however, the predictions differ between the CAM and the CX-COM model with $\alpha = 1$ (compare panels c and g).

Overall, the simulations demonstrate that CX-COM is able to account for important empirical findings in the judgment literature with the *alpha* parameter reflecting different levels of cue-abstraction processes[4] . Although these results suggests that CX-COM is more flexible than either the CAM or the exemplar model, it still provides a more parsimonious explanation than a blending model like Rulex-J (Bröder et al., 2017).

[4]Please note that a quantitative analysis of existing data in a more traditional paradigm does not allow to distinguish CX-COM from previously proposed judgment models. Due to a low number of observations per item the models are not recoverable. See Appending E for more details.

## Testing CX-COM's new predictions of judgment behavior

In the previous section we showed that CX-COMs can capture patterns of judgments reported in the literature and usually attributed to a shift in judgment strategies. However, the data of these studies does not allow comparing CX-COM with the other models because usually each item is only repeated once or twice making it impossible to distinguish the models. The reason is that CX-COM's unique characteristic is the shape of the response distribution and thus it only makes different predictions if an item is repeated many times. Accordingly, we conducted two new experiments to quantitatively and qualitatively test CX-COM's predictions.

**Quantitative test.** CX-COM combines the judgment processes of two very well established cognitive models, the CAM and the exemplar model. To quantitatively test CX-COM, we compared it against several competitors: an exemplar model, a CAM, RulEx-J, and a baseline model. The baseline model provides a benchmark for the absolute fit of the model. In the baseline model, we assume that participants respond with a constant value (with added noise), i.e. we fit a normal distribution with mean and variance as free parameters to participants. To better take CX-COM's complexity into account, we also did a cross validation for both experiments. All details concerning the fitting procedure and mean parameter values are shown in Appendix B.

**Qualitative test.** The CX-COM model predicts judgment patterns that are qualitatively distinct from single exemplar and cue-abstraction models. Specifically, CX-COM's competitive retrieval mechanism predicts multimodal response distributions and systematic changes in variability across items. This is in stark contrast to the classical exemplar models with integrative retrieval and the CAM which always predict a unimodal distribution centered around one model-predicted value. We tested the assumption of a competitive retrieval process explicitly in Experiment 2: A competitive retrieval process predicts that variations in judgments across and within items depend on the number of similar exemplars in memory and the distance between criterion values for similar exemplars. If a probe activates only one similar exemplar, the variability should be lower than if several similar exemplars are activated. If several exemplars with similar judgment values are activated, the variability in judgments is low. But if a probe activates exemplars with strongly dissimilar criterion values, high judgment variability and multimodal response distributions are predicted.

## Experiment 1

Experiment 1 was designed as a first, quantitative test for the CX-COM model. Specifically, we aimed at testing CX-COM in a situation that would usually favor an (integrative) exemplar model. In the experiment, participants had to solve a quantitative judgment task using three cues. To encourage exemplar retrieval, participants learned to judge a small set of training items and their criterion values

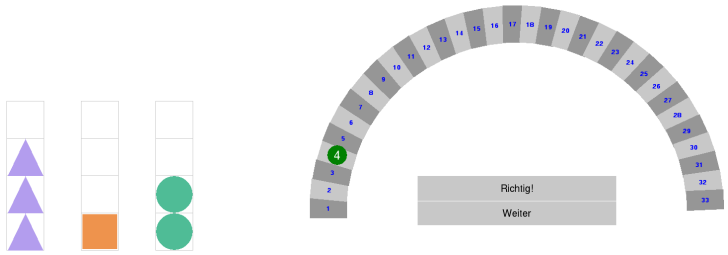


*Figure 4*. Visual presentation of stimuli in Experiment 1. The depicted stimuli 3.1.2 (on the left) is associated with criterion value 4 (highlighted on the half circle on the right).
*Note:* The text on the figure is German stating "Richtig!" for "Correct!" and "Weiter" for "Next."

by heart (Rouder & Ratcliff, 2006). After this training phase, they were instructed to repeatedly judge the criterion values of novel test items on the basis of their similarities to the training items.

## Method

***Participants.*** We tested 29 current or former students from the University of Basel ($M_{age}$ = 27 years, $SD$ = 7, range: 20–46 years). The target sample size was a priori set to 30 following conventions for one condition in cognitive modeling research (e.g. Hoffmann et al., 2016; Tsetsos et al., 2016). 35 participants were invited through the recruitment platform of the center for Economic Psychology in Basel and 29 came at the assigned time. The experiment took on average approximately 1 hour. Participants could choose between course credit or a payment of 20 Swiss francs per hour. In addition, participants could earn a performance-dependant bonus of 5 Swiss francs. The study received ethics approval by the Institutional Review Board (IRB) of the Faculty of Psychology at the University of Basel.

***Material.*** In Experiment 1 we used items with three dimensions shown as three adjoining stacks (left, middle, right) on the left side of a computer screen, see Figure 4. Each dimension was assigned a value between 1 and 4, indicated by the number of geometric shapes in the stack. Additionally, the dimensions differed in geometric shape (triangle, square, circle) and color (blue, red, green). Colors were chosen as complementary colors from the color wheel rendering them all similarly visually salient. Associated criterion values ranged between 1 and 33 and were presented on a half-circle on the right side of the computer screen. Figure 4 shows the visual presentation of stimuli as shown to the participants. Positions of the different cue dimensions were randomized across participants. Throughout the text, items are named with their three cue values separated by a dot. Item 3.1.1, for example, corresponds to an item with value 3 in cue dimension 1 (i.e. three shapes in the left position) and 1 in cue dimensions 2 and 3 (i.e. one shape in the middle and the right position).

To foster the use of exemplar-based processes we used a multiplicative environment (Hoffmann

et al., 2016; Juslin et al., 2008) and instructed participants to use similarities to judge novel items (Olsson, Enkvist, & Juslin, 2006). To help ensure that participants did not rely only on simple visual features of the items, for example, the higher the cue value the higher the criterion value, the first cue was inverted so that

$$j = (5 - c_1) \cdot c_2 \cdot c_3, \tag{12}$$

with cue values $c_1, ..., c_3$.

We presented only a small number of training items (six) that had to be learned by heart but used a larger number of novel test items (14) to evaluate participants' responses. The six training items were chosen such that the associated criterion values represented the general trend found in multiplicative environments; the lower part of the response scale was densely packed with observations (criterion values 4, 6, 8); in the higher part of the scale, single observations were rather sparse (criterion values 12, 18, 24).

Ten out of 14 test items were chosen so that each was most similar to one of the training items. The similarities are calculated with the city-block distance metric and assuming equal dimension weights. Four additional test items were included as fillers for which we did not systematically vary/control the similarity to all training items (see Table 2). We chose a relatively small number of training items and larger number of test items for two reasons: First, we wanted direct control over the similarity structure among training items and between training and test items. Second, we wanted participants to learn the training items by heart and to remember them throughout the whole experiment.

***Procedure.*** The experiment included three phases. In the training phase participants had to learn six different training items by heart. They were told that there was no simple functional dependency between the cues and criterion values and that they would later be asked to use the learned items to estimate the criterion values for novel items. We used two types of training blocks, judgment learning and cue learning. In the judgment-learning blocks participants were presented with the cue values of a training item on the left-hand side of the screen and had to choose the associated criterion value from the response circle on the right-hand side of the screen. In the cue-learning blocks one specific criterion value was highlighted on the response circle and participants were asked to adjust the stacks so that they represented the cues of the training item corresponding to the displayed criterion value. Participants could adjust the stack by repeatedly clicking on it to change the number of displayed shapes. After giving a response, participants received feedback in all trials during training. In total, the training phase consisted of 10 blocks.[5] Within each block the six training items were presented once in a random order. The training phase included six judgment-learning and four

[5]Two participants only completed 8 training blocks due to technical issues.

Table 2

*Stimuli, manipulation, and results in Experiment 1*

| | Training item | | | | | | Result |
|---|---|---|---|---|---|---|---|
| Criterion | 4 | 6 | 8 | 12 | 18 | 24 | |
| test item | 3.1.2 | 2.1.2 | 3.2.2 | 1.1.3 | 2.2.3 | 1.2.3 | Mean (*SD*) |
| 3.1.1 | **1** | 2 | 2 | 4 | 4 | 5 | 6.37 (4.13) |
| 4.1.2 | **1** | 2 | 2 | 4 | 4 | 5 | 8.53 (5.59) |
| 2.1.1 | 2 | **1** | 3 | 3 | 3 | 4 | 4.50 (2.79) |
| 3.2.1 | 2 | 3 | **1** | 5 | 3 | 4 | 10.93 (6.66) |
| 4.2.2 | 2 | 3 | **1** | 5 | 3 | 4 | 10.35 (4.68) |
| 1.1.4 | 4 | 3 | 5 | **1** | 3 | 2 | 18.49 (4.82) |
| 2.2.4 | 4 | 3 | 3 | 3 | **1** | 2 | 21.92 (3.71) |
| 2.3.3 | 4 | 3 | 3 | 3 | **1** | 2 | 19.78 (6.47) |
| 1.2.4 | 5 | 4 | 4 | 2 | 2 | **1** | 24.74 (2.96) |
| 1.3.3 | 5 | 4 | 4 | 2 | 2 | **1** | 21.27 (5.86) |
| 2.3.2 | 3 | 2 | 2 | 4 | 2 | 3 | 15.00 (5.98) |
| 2.1.3 | 2 | 1 | 3 | 1 | 1 | 2 | 12.61 (4.47) |
| 1.3.2 | 4 | 3 | 3 | 3 | 3 | 2 | 15.57 (5.54) |
| 2.3.1 | 4 | 3 | 3 | 5 | 3 | 4 | 14.77 (5.62) |

*Note.* Distance profiles for stimuli used in Experiment 1. Items consist of three cue dimensions (cf. Figure 4). Dimension values are separated by a dot. The distances are calculated using the city-block metric assuming attention weights of 1 in each dimension. Items with a distance of 1 to a test item are assumed to be most similar to that item and are shown in bold.

cue-learning blocks presented in alternating order, and with two judgment-learning blocks at the beginning and one judgment-learning block at the end of training. Participants received a bonus of 5 Swiss francs if they were able to correctly judge all training items in at least two subsequent training blocks.

In the test phase, participants saw 14 different novel test items and were asked to estimate the associated criterion values. People were asked to make the judgment according to the test items' similarities to the training items. The test phase included 15 test blocks. In each block all 14 items were presented in random order, resulting in 210 test trials. After test blocks 5 and 10, participants again judged the criterion values for all training items twice, resulting in four additional judgment-learning blocks during the test phase. These blocks were announced as training blocks and participants received feedback.

Table 3

*Model overview and results of the model comparison*

| Model | Experiment 1 | | | | Experiment 2 | | | | Model type |
|---|---|---|---|---|---|---|---|---|---|
| | Par. | Subj. | Deviance | BIC | Par. | Subj. | Deviance | BIC | |
| CAM | 5 | 7 | 1,294 | 1,321 | 6 | 9 | 1,077 | 1,109 | Cue abstraction |
| Exemplar | 4 | 0 | 1,375 | 1,397 | 5 | 2 | 1,119 | 1,146 | Exemplar |
| **CX-COM** | **5** | **18** | **1,268** | **1,294** | **6** | **19** | **1,033** | **1,065** | Mixture |
| RulEx-J | 6 | 4 | 1,285 | 1,317 | 7 | 1 | 1,075 | 1,112 | Mixture |
| Baseline | 2 | 0 | 1,490 | 1,501 | 2 | 0 | 1,400 | 1,411 | - |

*Note.* Par = Number of parameters; BIC = mean Bayesian information criterion; Subj = number of participants for whom a specific model had the best BIC-value. Best fitting model is shown in bold.

After completing the test phase, participants judged on a paper-and-pencil questionnaire how similar each test item was to each of the six training items, resulting in a total of 84 similarity judgments. Each page of the questionnaire showed one test item and the six training items. Participants were asked to judge the similarity on a scale of 0 (*completely different*) to 10 (*exactly the same*) for each pair.[6] The results of the similarity questionnaire are presented in Appendix A.

## Results

***Performance.*** In the last two training blocks (one cue-learning block and one judgment-learning block) participants judged 80% of the training items correctly. In the four judgment training blocks in the test phase they judged 81% of the training items correctly.

***Quantitative model evaluation.*** To get an idea whether competitive retrieval in general and CX-COM in particular explain the judgment behavior in experiment 1, we compared it with the exemplar model, the CAM, and the blending model RulEx-J. We also included a baseline model which assumes that participants respond the same, random value in every trial. We fit the five models to single-participant responses with a maximum likelihood estimation method. We used the BIC (Schwarz et al., 1978) to choose the best model for a single participant. A more detailed explanation about the fitting methodology, with an overview of best fitting parameter values, is presented in Appendix B. In addition we conducted a model comparison based on a cross validation reported in Appendix E.[7]

Across different model selection criteria, the exemplar model with competitive memory retrieval

[6]Two participants did not finish the similarity questionnaire and were excluded from the related statistics.

[7]The cross validation also suggested that CX-COM best described the largest number of participants, however its advantage was reduced and the Rules-J emerged as the second best model.

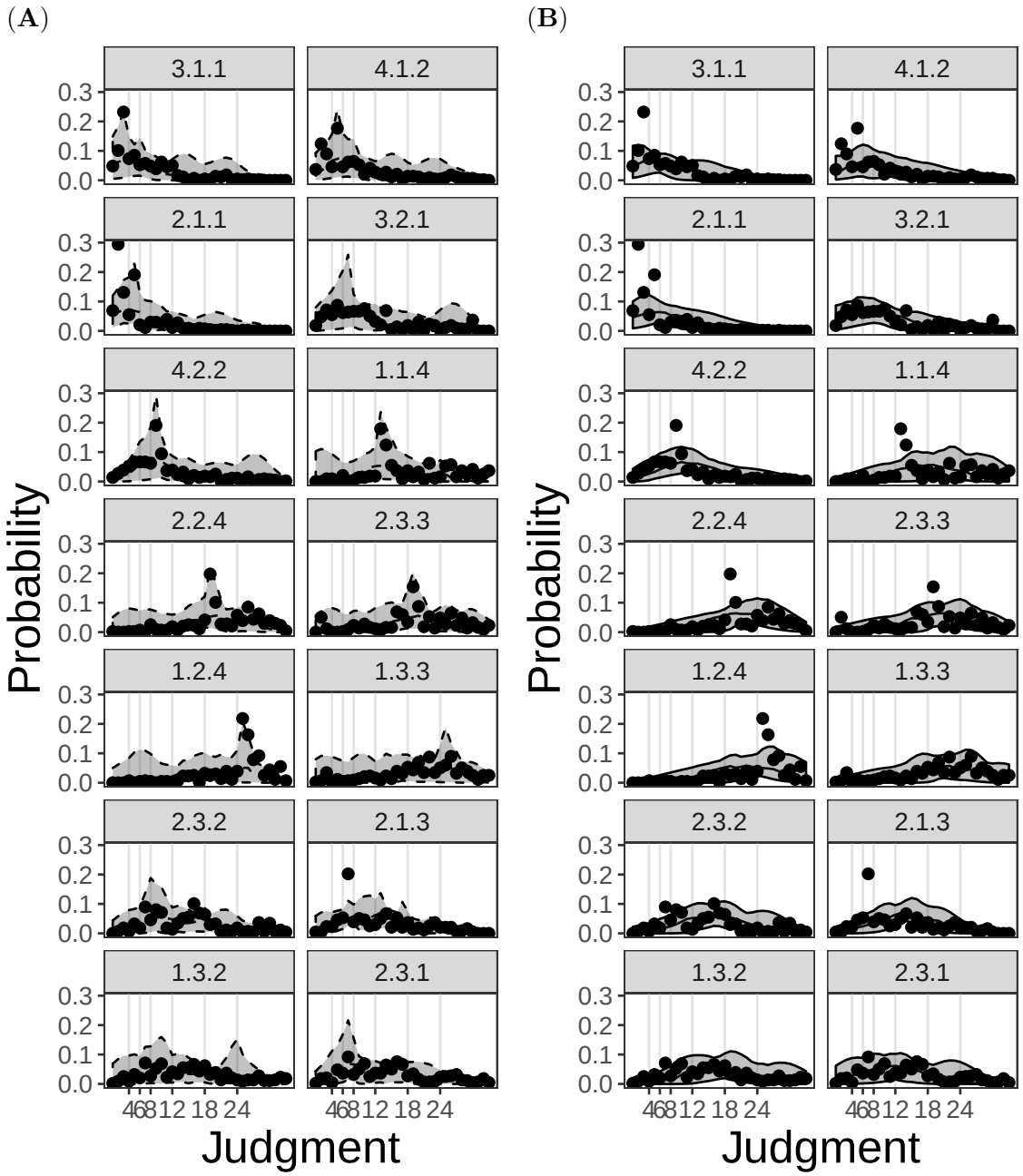


*Figure 5*. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions calculated across participants for all items in Experiment 1. Participants probability to respond a specific value (between 1 and 33) is shown as dots. Light grey lines indicate criterion values of the training exemplars.

(**A**) CX-COM-predictions (dashed line)

(**B**) CAM-predictions (solid line)

and a cue-abstraction component (CX-COM) fits the data very well. It is the model with the lowest mean BIC and deviance (Table 3) for over 60% of the participants and thus the most likely model to

describe their underlying cognitive process. The second best model is the cue-abstraction model (CAM) which is the most appropriate model for approximately 25% of participants followed by RulEx-J most appropriate for just under 15% of participants. The baseline model fares far worse than the other four models in relative and in absolute terms. The average deviance for the baseline model is 200 points higher than for the other models. However, it is not always the worst model. In fact, CAM is worse than the baseline model for one participant

A model recovery based on the design of Experiment 1 confirmed the ability of CX-COM and the CAM to discriminate when participants used these different cognitive process. When CX-COM generated the data then 78% of the time it was identified as the data-generating model. When the CAM generated the data then 96% of the time it was identified as the data-generating model. Note that CX-COM can have overlapping predictions with the CAM, depending on the values of the attention weights. We designed Experiment 2 to discern the two models also qualitatively.

The quantitative differences in model fits are also illustrated in Figure 5 comparing participants' response distributions for each test item with the predictions of CX-COM (Figure 5A) and the second best model CAM (Figure 5B). Most test items for which a very similar training item exists (e.g. items 3.1.1 to 1.3.3 in the figures, for reference see Table 2) show a clear peak in participants response probability for a judgment close to the value associated with the very similar training item. CX-COM is able to predict these peaks nicely while the CAM and the exemplar model cannot.

However, the BIC does not adequately reflect the functional flexibility of the models, in particular Rulex-J and CX-COM, so that we additionally conducted a cross-validation study. The details of the cross validation procedure are described in Appendix B and the results are shown in Table B1. Compared to the model selection based on BIC, CX-COM still is the best model for most participants and describes 14 participants best (4 participants less than previously). The mixture model RulEx-J explains 13 participants best (9 participants more than according to BIC) and CAM only explains 2 participants best (5 participants less). Somewhat surprisingly, the results show a larger advantage for the models with a high functional complexity, that is RuleEx-J and CX-COM, compared to models with a lower functional complexity (the CAM and the exemplar model). However, cross-validation methods have been shown to favor more complex models (Browne, 2000), which might explain why Rulex-J, the arguably most complex model, had the largest gain. Overall, the result suggests that the additional functional complexity introduced by mixture models is warranted in our task.

## Discussion

In Experiment 1 we tested the CX-COM model quantitatively against competing models from the literature. CX-COM assumes that judgments are the result of (1) a competitive retrieval

mechanism and (2) a subsequent cue-abstraction mechanism that adjusts the criterion value of the recalled exemplar. We compared CX-COM to an exemplar model, a cue-abstraction model (CAM) and the blending model RulEx-J.

CX-COM captured the judgments best compared to the competing models in terms of number of assigned participants according to BIC, mean BIC, and mean deviance. It is still the best model for most participants according to the cross validation. As Figure 5 shows, CX-COM is able to capture almost every peak in the participants' probability distribution, while the CAM, the second best model according to the BIC, cannot. Interestingly, there is no participant assigned to the (pure) exemplar model. That is, even in a multiplicative environment that is usually understood to promote the use of an exemplar process (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012), we found evidence that all participants used a cue-abstraction process. This result suggests that adjustments based on beliefs about the cue–criterion relations play an important part in judgments, even when overall judgments may be best described by an exemplar model. To replicate the quantitative results and to qualitatively test the assumption that previously encountered exemplars compete for retrieval we designed Experiment 2.

## Experiment 2

In Experiment 2 we focused on the prediction that sets CX-COM apart from other models in the literature: The assumption that exemplar retrieval is competitive and not integrative. For integrative retrieval, the judgment is based on the similarity to all previously encountered instances, independent of how many exemplars are similar to the probe and how strongly their criterion values differ from one another. Therefore, an integrative retrieval component predicts unimodal response distributions and no systematic variation in judgments across items.

In contrast, during competitive retrieval one exemplar is recalled on each retrieval attempt, implying that judgments for each probe vary depending on how similar (or dissimilar) the criterion values for similar exemplars in memory are. If only one exemplar in memory is highly similar to the probe, retrieval probability for this exemplar is high and it is recalled most of the time. Thus judgments for this probe will vary only to a small extent between trials. However, if two exemplars are highly similar to the probe, both exemplars are, in principle, recalled equally often. Response variability depends on the distance between the associated criterion values. If the distance is small, say, the decision maker retrieves criterion values of 25 and 33 on a scale of 1 to 33, judgments for this probe should vary little across trials. If the distance between the associated criterion values is large, for instance, 9 and 33, judgments should vary strongly. In addition, the distribution of judgments should be bimodal, within and across participants. We tested this prediction by systematically manipulating the number of exemplars with a high recall probability and the distance between associated criterion

(**A**) (**B**)

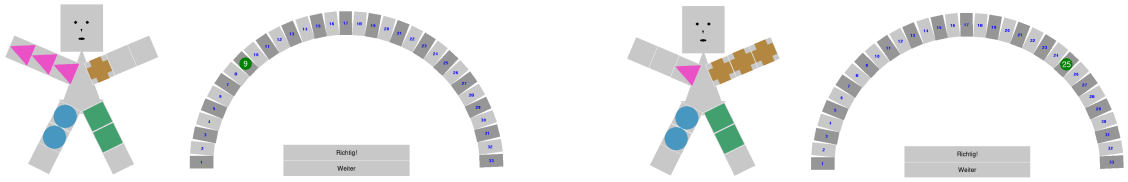

*Figure 6*. Example stimuli used in Experiment 2. The text on the figure is German stating "Richtig!" for "Correct!" and "Weiter" for "Next."
(**A**) Stimulus 3.2.2.1; high value in a left extremity is associated with a low value on the scale (value 9, left part of the scale).
(**B**) Stimulus 1.2.2.3; high value in a right extremity is associated with a high value on the scale (value 25, right part of the scale).

values. Recall probabilities depend on the similarity between a test item and exemplars in memory. Response variability depends on the interplay between the perceived difference in similarity and the distance in the associated criterion values.

## Method

***Participants.*** We tested 33 current or former students from the University of Basel ($M_{age}$ = 28 years, $SD$ = 8, range: 19–48 years).[8] The target sample size was a priori set to 30 following conventions for one condition in cognitive modeling research (e.g. Hoffmann et al., 2016; Tsetsos et al., 2016). 35 participants were invited through the recruitment platform of the center for Economic Psychology in Basel and 33 came at the assigned time. The experiment took on average 1 hour. Participants received course credit or an hourly payment of 20 Swiss francs. In addition, participants could earn a bonus of up to 5 Swiss francs. The study received ethics approval by the Institutional Review Board (IRB) of the Faculty of Psychology at the University of Basel.

***Material.*** The general setup in Experiment 2 was very similar to the setup from Experiment 1. In Experiment 2 we used stimuli with four cue dimensions and three possible values on each dimension (see Figure 6). Each cue dimension was represented by a limb of a robot. Each limb had a certain number of slots for power modules (the cue values) and was associated with a different geometric form (triangle, square, circle, cross) and color (red, blue, green, brown). Participants were told to judge the overall power level of the robot (the criterion with the response scale again between 1

[8] Participant information for three participants are missing due to technical problems.

and 33) and that the power level depended on the number of power modules in the limbs. Figure 6 shows the training stimuli as they were presented to the participants. Positions of the different cue dimensions were partly randomized.

The overall power level was a linear function of the number of power modules in each limb,

$$j = -15 + 4 \cdot c_1 + 12 \cdot c_4, \tag{13}$$

with cue values $c_1, ..., c_4$. The second and the third cue were not predictive of the response.

Because the predictive cues were positively related to the criterion, participants should have been able to rapidly identify the cue–criterion relationship (see Table 4). Small values on all dimensions indicated small criterion values and large cue values indicated large criterion values respectively. However, different values on cue dimensions 1 and 4 required additional attention. A large value on cue dimension 4 and a small value on cue dimension 1 indicated a large criterion value, whereas a large value on cue dimension 1 and a small value on cue dimension 4 indicated a small criterion value (see Table 4 and for an example, Figure 6).

The simple rule underlying stimulus generation allowed us to manipulate the distance between criterion values associated with training items with high recall probabilities. Thus, if the cue value of a test item on dimension 1 was small but all other cue values were large, then a judgment based on a similar training item should have been very likely to be large as well (see test item 1.3.3.3 in Table 4). However, if the cue value of a test item on dimension 4 was small but all other cue values were large, a judgment based on a similar training item might be large or small (see test item 3.3.3.1 in Table 4).

To extend the manipulation of distance between most similar exemplars conceptually to all items, we developed a measure we call the similarity neighborhood (SN, see Table 4). The SN score for a test item was calculated as the mean distance between the criterion values for those training items most similar to the test item (assuming city-block distance). The higher the mean distance is, the higher the expected variability in judgments. For example, the two training items which are most similar to test item 3.3.3.1 are items 3.2.2.1 (criterion value 9) and 3.3.3.3 (criterion value 33). The distance between the criterion values associated with the training items—the SN score—is 24 (Table 4). Considering item 1.3.3.3, the most similar items are 1.2.2.3 (criterion value 25) and 3.3.3.3 (criterion value 33) with a distance between criterion values of only 8. For items with only one most similar training item (e.g. 3.3.3.2 and 2.3.3.3) we used the mean distance between criterion values of the most and second most similar training items as SN score. Item 2.3.3.3, for example, is most similar to item 3.3.3.3 and second most similar to items 1.2.2.3 (with distance between criterion values of 8) and 2.3.1.2 (with distance between criterion values of 16). The SN score is, thus, 12. With this approach we considered retrieval candidates with a combined recall probability of over 90% per item ($M = 0.97, SD = 0.03$).

Table 4

*Stimuli and results in Experiment 2*

| | Training item | | | | | SN | Results | |
|---|---|---|---|---|---|---|---|---|
| Judgment | 1 | 9 | 17 | 25 | 33 | | | |
| test item | 1.1.1.1 | 3.2.2.1 | 2.3.1.2 | 1.2.2.3 | 3.3.3.3 | | Mean (*SD*) | *SD* (*SD*) |
| 1.2.2.1 | **2** | **2** | 4 | **2** | 6 | 16 | 10.08 (2.93) | 3.11 (1.78) |
| 3.2.2.3 | 6 | **2** | 4 | **2** | **2** | 16 | 26.39 (2.52) | 2.77 (1.96) |
| 2.2.2.2 | 4 | **2** | **2** | **2** | 4 | 11 | 18.22 (2.76) | 3.49 (2.07) |
| 3.3.3.1 | 6 | **2** | 4 | 6 | **2** | 24 | 18.04 (5.32) | 4.63 (2.51) |
| 1.3.3.3 | 6 | 6 | 4 | **2** | **2** | 8 | 27.34 (2.28) | 2.62 (1.91) |
| 1.1.1.3 | **2** | 6 | 4 | **2** | 6 | 24 | 17.52 (5.48) | 3.33 (2.04) |
| 3.1.1.1 | **2** | **2** | 4 | 6 | 6 | 8 | 7.93 (2.92) | 2.49 (2.11) |
| 2.1.1.1 | **1** | 3 | 3 | 5 | 7 | 12 | 6.48 (2.94) | 2.5 (1.86) |
| 1.1.1.2 | **1** | 5 | 3 | 3 | 7 | 20 | 11.3 (4.34) | 3.61 (1.95) |
| 2.3.3.3 | 7 | 5 | 3 | 3 | **1** | 12 | 29.87 (2.09) | 2.25 (2.24) |
| 3.3.3.2 | 7 | 3 | 3 | 5 | **1** | 20 | 25.59 (4.35) | 4.11 (2.78) |
| 2.2.2.1 | 3 | **1** | 3 | 3 | 5 | 11 | 11.22 (3.22) | 3.04 (1.52) |
| 3.3.2.1 | 5 | **1** | 5 | 5 | 3 | 24 | 13.98 (4.05) | 3.38 (2.14) |
| 1.2.2.2 | 3 | 3 | 3 | **1** | 5 | 16 | 16.84 (3.09) | 3.7 (2.05) |
| 1.2.3.3 | 5 | 5 | 3 | **1** | 3 | 8 | 24.78 (2.97) | 2.62 (1.53) |
| 2.2.1.2 | 3 | 3 | **1** | 3 | 5 | 11 | 15.39 (2.81) | 3.3 (2.07) |
| 2.3.2.2 | 5 | 3 | **1** | 3 | 3 | 11 | 19.99 (2.86) | 3.24 (2.04) |
| 3.2.2.1 | 4 | **0** | 4 | 4 | 4 | 14 | 11.62 (3.74) | 2.31 (1.65) |
| 2.3.1.2 | 4 | 4 | **0** | 4 | 4 | 12 | 17.47 (2.24) | 2.58 (1.82) |
| 1.2.2.3 | 4 | 4 | 4 | **0** | 4 | 14 | 23.2 (2.42) | 2.94 (2.4) |

*Note.* Distance profiles for stimuli used in Experiment 2. Items consist of four cue dimensions (cf. Figure 6). Dimension values are separated by a dot. The distances are calculated using the city-block metric and all attention weights set to 1. Items with the lowest distance are shown in bold. SN (similarity neighborhood) is the mean distance of the criterion values for similar training items. If more than one training item is most similar to a test item (first seven items), SN is the mean distance among the criterion values of these training items. If only one training item is most similar to a test item, then SN is the mean distance between the most similar and second most similar training item. Results (mean and *SD*) are calculated within participants and then averaged across participants.

***Procedure.*** The experiment consisted of two phases. In the training phase participants had to learn five different items by heart. They were told that they would later be asked to use these items to estimate the criterion value for novel items. During training, participants were presented with a training item on the left-hand side of the screen and had to choose the associated criterion value from the response circle on the right-hand side of the screen. Participants received feedback in all trials during training.

The maximum number of training blocks was set to 12. However, participants with 100% accuracy in three blocks (with 100% accuracy in at least two subsequent blocks) had to complete only one more block to move on to the test phase. People who failed to reach this criterion had to complete all 12 training blocks and then moved on to the test phase.

In the test phase, participants had to judge 20 different items without feedback. Seventeen items were unknown and three were training items (3.2.2.1, 2.1.3.2, 1.2.2.3). Participants were asked to estimate the criterion values on the basis of their similarity to the items learned during training. The test phase included 10 test blocks. In each block all 20 items were presented in a randomized order, resulting in a total of 200 test trials. Participants received a bonus relative to their accuracy compared to an integrative exemplar model without cue abstraction (assuming equal weights) during the test phase. The maximal bonus was set to 5 Swiss francs.

## Results

***Performance.*** On average, participants completed training successfully after 7.9 blocks ($SD = 2.3$). Two participants failed to reach the inclusion criterion and were excluded from further analyses.

***Multimodality and across-item variability.*** In this experiment we contrasted the competitive and integrative retrieval mechanism. Items were chosen to test two key predictions of competitive retrieval: The occurrence of multimodal response distributions within and across participants and across-item variability being a function of the similarity structure of learned exemplars. A list of all items is shown in Table 4. Recall that we hypothesized (a) that the number of training items that are very similar to a test item and their absolute difference in criterion values influences the test item's variability (e.g. items 3.3.3.2 and 1.3.3.3 should have a lower variability than item 3.3.3.1), and (b) that we would be able to observe a response distribution with a visible bimodal shape in test items that are very similar to two training items that have a high absolute distance between their criterion values (e.g. item 3.3.3.1).

*Multimodality: Descriptive statistics.* Figure 7 shows response distributions for items 3.3.3.1 and 1.3.3.3. The figure includes two test items that are similar to only one training item but have similar cue values (3.3.3.2 and 2.3.3.3) as well as two training items tested again during the test (a

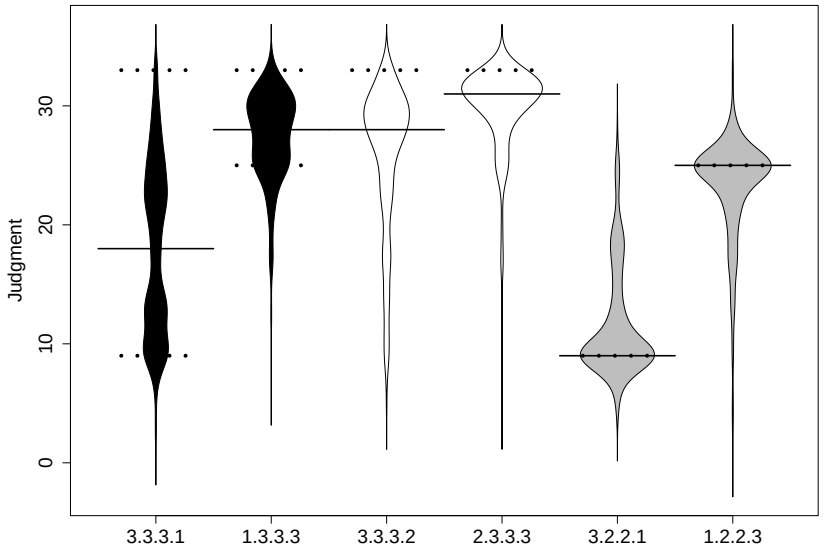


*Figure 7.* Participants' response distributions for items representative of the manipulation in Experiment 2. Black distributions correspond to test items with two most similar training items but a high (3.3.3.1) and low (1.3.3.3) distance between associated criterion values. White distributions correspond to test items with one most similar training item. Response distributions for training items judged at test are displayed in gray. Thick dotted lines correspond to criterion value of maximally similar training items; thin lines show the median of the distribution. The standard deviation of the kernel estimation was set to 1.06 (Scott, 1992).

beanplot including all tested items is shown in Appendix C). For the training items, the median judgment during test corresponds to the learned criterion value and the standard deviation of responses is low. In items 3.3.3.2 and 2.3.3.3, the standard deviation is lower than in items 3.3.3.1 and 1.3.3.3. Additionally, the response distribution has a different shape. Especially, item 3.3.3.1 clearly shows a multimodal response distribution with the most frequent responses close to the learned criterion values of the two most similar training items. The deviation of the most frequent responses from the learned criterion values is consistent with cue-based adjustment.

Response distributions aggregated across participants were multimodal for items with a clear effect of similarity and distance, for example, item 3.3.3.1 in Figure 7. As an additional piece of evidence to corroborate our claim that this is the result of a competitive retrieval and not, for example, the result of variability in parameter settings or strategies between participants, we show model-predicted response distributions and observations for the two participants who were best fit by CX-COM (Figure 8) and the CAM (Figure 9), relative to the respective other model according to the difference in BIC. The participant best fit by CX-COM (Figure 8) displays multimodal response

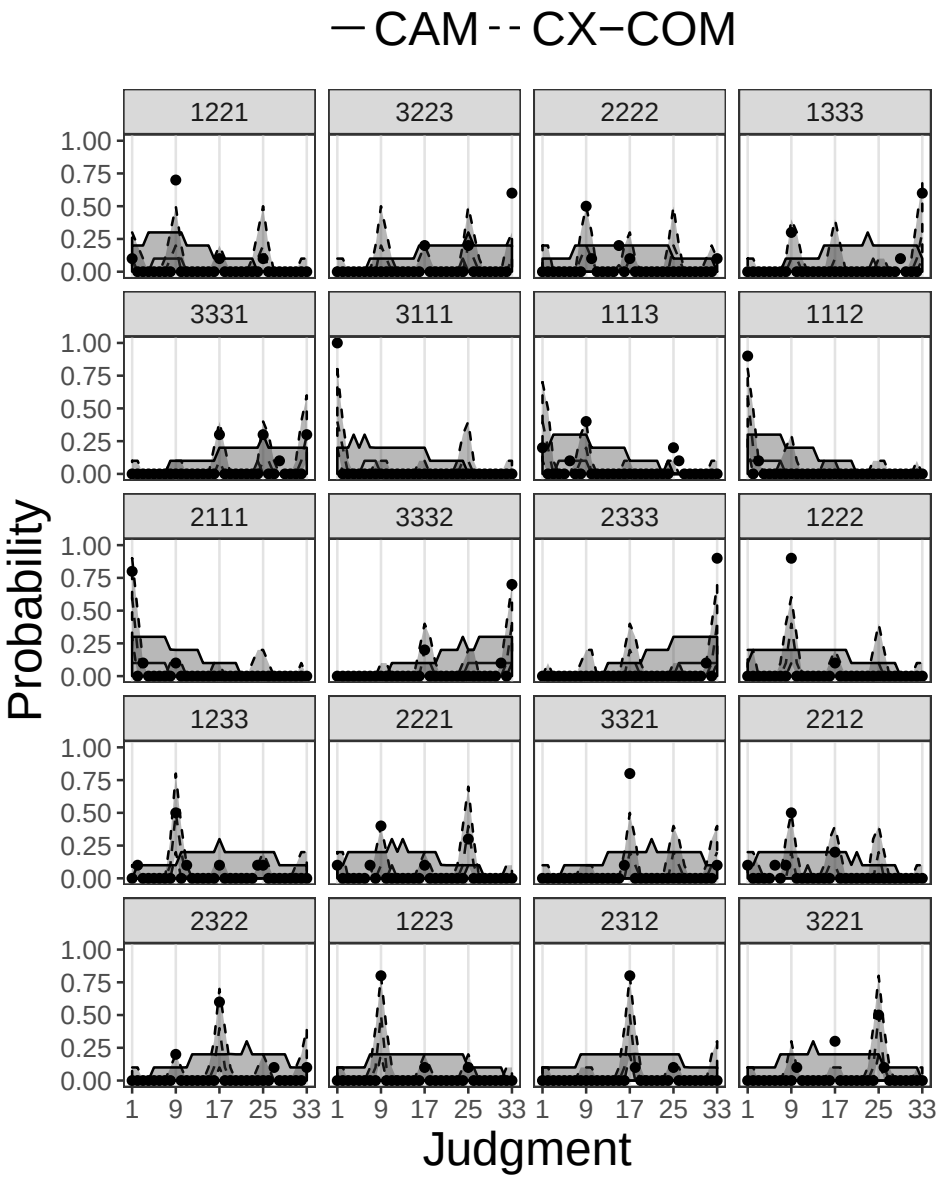


*Figure 8*. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions for the participant best described by CX-COM relative to the CAM (i.e. highest difference in BICs). The participant best described by the CAM is shown in Figure 9. Light grey lines indicate criterion values of the training exemplars.

distributions as predicted by CX-COM. CX-COM captures the responses of the participant well, the broad distributions predicted by the CAM do not.

*Multimodality: Inferential statistics.* In order to substantiate our claim and the descriptive results, we tested for multimodality across participants using Hartigan's dip test (Hartigan & Hartigan, 1985). Conceptually, Hartigan's dip test calculates the maximum distance between an empirical distribution and the best fitting unimodal distribution. In principle, all response distributions predicted by CX-COM are multimodal, the predicted response distribution can have as many modes as

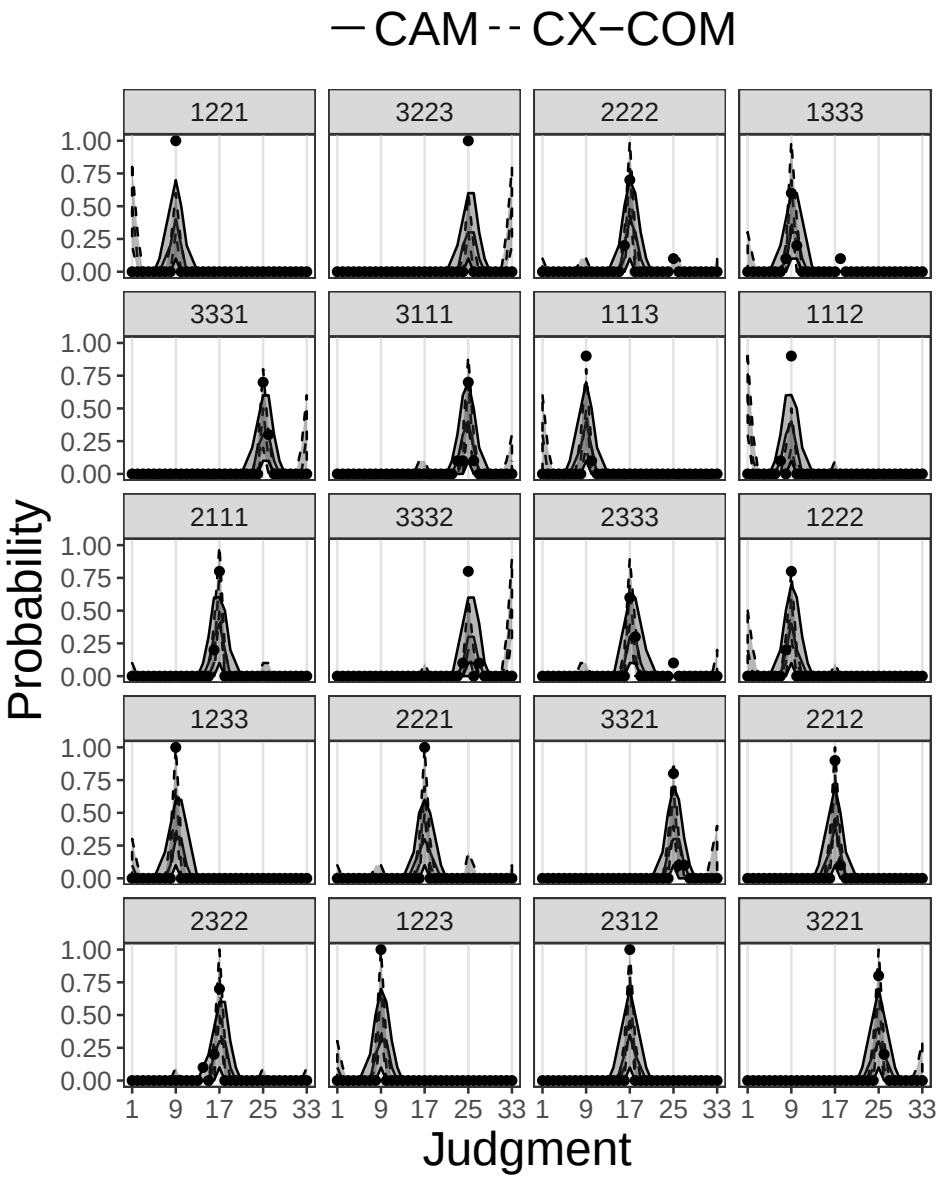


*Figure 9*. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions for the participant best described by the CAM relative to CX-COM (i.e. highest difference in BICs). The participant best described by CX-COM is shown in Figure 8. Light grey lines indicate criterion values of the training exemplars.

there are learned exemplars (cf. Figure 1). However, it can be very hard to detect this multimodality, for example, if the recall probability for one exemplar is very high. In this case, the multimodal structure of the distribution is easily confused with a distribution with long tails and a substantial number of observations are needed to classify the distribution correctly. Aggregated across participants, 17 out of 20 test items showed significant multimodality (mean $D = .06, p < .01$, bonferroni corrected). The 3 test items which showed no signs of mulitmodality are the three training items repeated during the test phase (mean $D = 0.03$). For participants best fit by the CX-COM

model (19 participants) 14 out of 20 items showed significant multimodality and for participants best fit by the CAM (9 participants) only 4 out of 20 items showed significant multimodality.

*Across-item variability.* Within participants we lacked the power to detect multimodal response distributions (see Appendix D for a post-hoc power analysis). We thus tested a second hypothesis: Across-item variability is a function of the distance between criterion values of highly similar training items. The higher the distance between the criterion values of two similar training items, the more variable should responses to this test item be.

To investigate the influence of competitive exemplar retrieval on judgment variability systematically across all test items, we predicted the variability in judgments with the SN score, that is the average distance between the criterion values of similar training items. To measure variability of judgments for an item we used the standard deviation of the judgments for an item during the test phase (calculated within participant and averaged across participants), see Table 4. On average, the SN score correlated positively with the mean standard deviation ($r = 0.66, p < 0.01$). Because we aimed at an analysis on the participant level, we used a linear mixed effects model that predicted an item's judgment variability with its SN score as a fixed factor and items' and participants' intercepts as random factors. This model predicted the items' judgment variability significantly better than a baseline model including only participants' and items' random intercepts, $\chi^2(1) = 10.80, p < 0.01$.

***Quantitative model evaluation.*** We fit the same models under the same conditions as in Experiment 1[9]. All models had one more free parameter because of the additional cue dimension (the respective dimension weight), except the baseline model that assumes a constant response. A detailed description of the fitted parameters and best fitting parameter values is given in Appendix B. A model comparison based on a cross validation is reported in Appendix E.

The CX-COM model again captured participants' judgments best. It was the most appropriate model for over 60% of the participants and had the lowest mean BIC (Table 3). Around 30% of participants were again best described by the CAM. Surprisingly, the blending model RulEx-J fared worse than in experiment 1 and described only one participant best. The baseline model again fit participant responses much worse than all other models.

[9]Following the suggestions of anonymous reviewers we also fitted a model with competitive retrieval and without a cue-abstraction component and two possible versions of a prototype model to the data in this experiment. In one version of the prototype model we assume the most extreme exemplars to be prototypes, i.e. 3.3.3.3 and 1.1.1.1 (Prot1). In the second version we assume the most informative exemplars to be prototypes, i.e. 1.2.2.3 and 3.2.2.1 (Prot2). None of these three models explain any participants best and the average BIC is much higher ($BIC_{Prot1} = 1139, BIC_{Prot2} = 1280, BIC_{Competitivew/ocue-abstraction} = 1220$) than the average BIC of the CAM ($BIC = 1109$) or the CX-COM model ($BIC = 1065$).

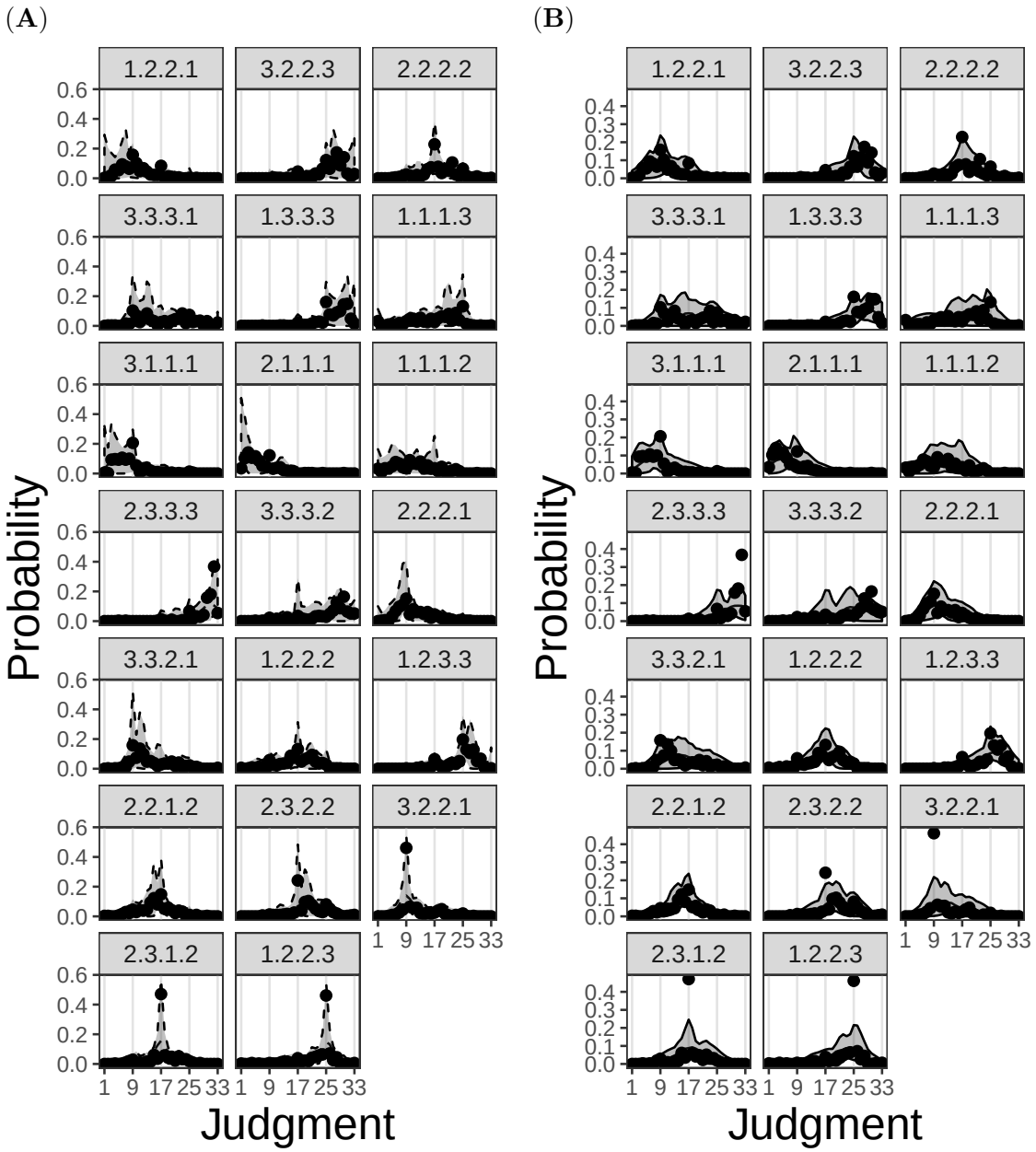


*Figure 10*. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions calculated across participants for all items in Experiment 2. Participants probability to respond a specific value (between 1 and 33) is shown as dots. Light grey lines indicate criterion values of the training exemplars.

(**A**) shows CX-COM-predictions (dashed line)

(**B**) shows CAM-predictions (solid line)

A model recovery with CX-COM and the CAM based on the design of Experiment 2 again supports these conclusions. When CX-COM generated the data, 94% of the time CX-COM was identified as the data-generating model. When the CAM generated the data, 87% of the time the CAM

was identified as the data-generating model. In the cross validation, CX-COM also emerged as the best model, followed by Rulex-J (see Appendix E).

Figure 10 shows a comparison of participants' aggregated response probabilities with aggregated predictions of CX-COM (Figure 10A) and the CAM (Figure 10B). Both models describe the aggregated responses well. However, CX-COM clearly captures some peaks that the CAM cannot.

We again conducted a cross validation in Experiment 2. The results are very similar to the results according to BIC (see Table B1). As in Experiment 1, the two mixture models CX-COM and RulEx-J explain the responses of most participants best despite their higher functional complexity. CX-COM is still the best model for 17 participants (only describes the judgments of 2 participants less), whereas RulEx-J explains the judgments of 11 participants best (10 participants more than according to BIC). The number of participants best explained by the CAM drops from 9 to 3. The exemplar model looses all its participants. These results again suggest that the additional complexity introduced by mixture models is warranted also according to Experiment 2.

## Discussion

Experiment 2 investigated whether an exemplar model with a competitive retrieval mechanism explains judgment behavior better than the traditional exemplar model using integrative retrieval. We found that the exemplar model with an integrative retrieval mechanism could neither quantitatively nor qualitatively account for the data. Test items varied in their trial-to-trial judgment variability and most items showed clear signs of multimodality. Furthermore, judgment variability across items was a function of the distance between the criterion values associated with similar training items: A prediction that cannot be accounted for by an integrative retrieval component or a CAM. The multimodal response patterns occurred across and more importantly within participants. The existence of multimodal response distributions within participants delivers important evidence in favor of exemplar models assuming a competitive retrieval mechanism. If multimodality was found only across participants, differences in parameter settings such as attention weights or the strategies used could also explain the results.

Quantitatively, the best model was again the CX-COM model that was most appropriate for over 60% of the participants in the model comparison based on BIC. The CAM also described some participants well. One possible reason is the linear structure of the task. This type of task is known to coincide with cue-abstraction strategies (Hoffmann et al., 2016; Olsson et al., 2006). However, the CAM also assumes that judgment variability is constant across items, an assumption that was clearly violated by the majority of participants.

The results from the cross validation likewise support CX-COM. The high number of participants best described by CX-COM and Rulex-J suggests that the majority of participants seems

to rely on both cue-abstraction and exemplar processes. Still, CX-COM clearly best captured the responses of more participants than Rulex-J. One reason for this could be that we designed it as a critical test for the prediction of multimodal responses. In sum, the results provide important evidence for competitive retrieval.

## General Discussion

When people evaluate objects and situations in order to form a decision or category assignment, research suggests that knowledge of previous experiences is combined with more abstract knowledge about a specific context (e.g. Erickson & Kruschke, 1998; Juslin et al., 2008). Judgment research has largely been mute about the concrete nature of the retrieval processes and possible combinations of recalled and abstracted knowledge. The present research sought to address these two shortcomings by spelling out a new cognitive model, CX-COM. Building on established models for quantitative judgments, CX-COM introduces a competitive retrieval mechanism to describe how exemplars are activated in memory and adjusts the judgment based on the retrieved exemplar using abstracted cue knowledge. To contrast its underpinning assumptions with competing theoretical ideas, we (a) tested CX-COM quantitatively against several competitor models from the literature and (b) derived and tested a qualitative prediction about the variability and shape of response distributions induced by CX-COM's competitive retrieval mechanism. Overall, the CX-COM model was best suited to explain human judgment behavior across the two experiments. Quantitatively, the model was most appropriate for describing the data of the majority of participants and also had on average the lowest BIC values. In addition, the qualitative test supported the model's assumptions that past exemplars compete for retrieval when people make judgments (Experiment 2). We next reconsider the model's assumptions in detail and then compare the mechanisms to similar theories in judgment, categorization, and function learning research.

### Competitive retrieval from exemplar memory

Traditionally, exemplar models in judgment and categorization have proposed that people retrieve a composite of all previously encountered exemplars from memory. This composite does not change across trials and a constant error is assumed. This implies that judgment variability is constant across items. In contrast, in both experiments we found evidence that judgment variability systematically varied across items, indicating competitive retrieval. A stricter test of this assumption in Experiment 2 suggested that some items elicited multimodal response distributions—across and within participants. Furthermore, across-item variability was a function of an item's similarity structure, consistent with the qualitative predictions of the CX-COM model. In line with this, more participants were best described by models assuming competitive retrieval from memory in both experiments.

Taken together, these results suggest that exemplar retrieval in quantitative judgments is best described by competitive retrieval processes, corresponding to established theories on retrieval processes in episodic memory (Anderson, 1983; Logan, 2002; Ratcliff, 1978) and process-oriented versions of exemplar models (Nosofsky & Palmeri, 1997; Palmeri, 1997). The response distributions are not consistent with the assumption that judgment variability is constant across items, which is the assumption made by exemplar models with an integrative memory component, the pure CAMs, as well as blending models such as Rulex-J.

Importantly, these results also highlight that the form of the response distribution and the variability of responses provide a tool to understand the nature of the involved cognitive processes (Kalish, Lewandowsky, & Kruschke, 2004). Moreover, the ability to explain and predict the expected variance in judgments is also of practical relevance given that it puts natural constraints on the expected reliability in judgments that will vary depending on the experiences of the decision maker.

## Combining cue abstraction and exemplar retrieval

Although judgment research has investigated how people shift between exemplar retrieval and abstracted knowledge, little empirical work has studied the degree to which the two processes are intertwined. Within CX-COM, we assumed that cue abstraction acts on the retrieval of stored exemplars. In both experiments, CX-COM consistently outperformed an exemplar model that did not consider any cue abstraction, independently of the tested environment. This suggests that beliefs about how cues are related to the criterion influence judgments in addition to memories of similar exemplars.

Besides supporting the idea that cue-abstraction processes exist in quantitative judgments, the two experiments also provided evidence for the effects of specific exemplars. Even in the linear judgment task in Experiment 2, CX-COM described participants' judgments better than the pure CAM, although a host of research suggests that in linear tasks, judgments are usually best described by the CAMs (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012). Furthermore, the multimodal response distributions follow naturally from the assumption of exemplar competition but cannot be explained by pure cue-abstraction processes.

Taken together, these results suggest that quantitative judgments are based on a combination of exemplars retrieved from memory and abstracted beliefs about the cues. They resonate well with previous empirical research showing that specific exemplars and rules simultaneously influence judgments and categorizations (Brooks & Hannah, 2006; Hahn et al., 2010; von Helversen et al., 2014) and research showing the advantage of mixture models in categorization (Erickson & Kruschke, 1998; Nosofsky et al., 1994; Vanpaemel & Storms, 2008) and function learning (DeLosh, Busemeyer, & McDaniel, 1997; Kalish et al., 2004).

## Relation to different approaches

The two experiments provide consistent support for the CX-COM model. This new model explains how beliefs about cue–criterion relationships interact with memories about specific instances. In the following we spell out similarities and differences between CX-COM and other models and approaches in related domains.

***Blending Models.*** Blending models are based on the assumption that an exemplar and a cue-abstraction component processes information independently and the response is a weighted average of the two results. The measurement model RulEx-J (Bröder et al., 2017) is a very recent and successful implementation of this idea in the domain of multiple-cue judgments. In line with findings in multiple-cue judgment, the mixture parameter in RulEx-J weighs the contribution of the two model components and reflects the employed strategy on the individual level and the impact of the environment or experimental instructions on the aggregate level.

In our two experiments the environments differed; In Experiment 1 we used a multiplicative environment that is known to coincide with exemplar processing and in experiment 2 we used a linear environment that is known to coincide with cue-abstraction processes. In line with findings from the literature and especially with the results presented by Bröder et al. (2017) we find on average a higher mixture parameter ($\beta$) in experiment 2 (.6) than in Experiment 1 (.5, see Table B1). These results suggest RulEx-J reflects differences in the amount of cue abstraction processes well. However, our CX-COM model outperformed RulEx-J consistently in the quantitative model comparison in both experiments. Additionally, RulEx-J is not able to account for multimodal response distributions and changes in variability across items.

***Function Learning.*** In both function learning and multiple-cue judgment, a numerical criterion has to be estimated given contextual information. However, function learning and multiple-cue judgment differ strongly in the complexity of the to-be-judged objects. In multiple-cue judgment the evaluation of objects is based on several cues with several possible values, while in function learning it is based on one numeric value. Accordingly, models from the function learning literature are rarely considered in the literature on multiple-cue judgment.

Function-learning research found that participants often extrapolated in a rule-based fashion, although they learned with single exemplars (DeLosh et al., 1997). Accordingly, theories in function learning often consider a competition between memory items (or rules) as well as mixtures between retrieval-based and cue-abstraction processes. Most notably, CX-COM could be considered as an extension of EXAM (DeLosh et al., 1997; McDaniel & Busemeyer, 2005) for judgments based on multiple cues. EXAM learns similarity-based associations between one-dimensional, quantitative inputs and outcomes. When generalizing to new patterns, it uses the distance between similar inputs to recruit a linear extrapolation mechanism. Thus, EXAM and CX-COM share the idea that a

cue-abstraction mechanism adjusts the response values of a recalled response. The difference between CX-COM and EXAM mirror the different complexities of the to-be-judged objects. In the EXAM model, the cue-abstraction component considers not only the one recalled output but also two outputs with similar input values. The response is based on the proportion of change in input and output values. In contrast, CX-COM adjusts the retrieved criterion depending on the difference in cue values between the probe and the one recalled exemplar.

***Knowledge Partitioning.*** In function learning and categorization, knowledge partitioning spells out the idea that knowledge is separated into independent parcels that potentially contain mutually contradictory information (Kalish et al., 2004; Lewandowsky, Roberts, & Yang, 2006). As a result of knowledge being spread out over a space of, potentially numerical, response values in separate parcels, knowledge partitioning also predicts multimodality of responses and these patterns have been found in the domain of function learning (Kalish et al., 2004).

The most prominent model implementing knowledge partitioning in the function-learning domain is POLE (Population Of Linear Experts; Kalish et al., 2004). According to POLE, judgments are based on linear experts, that is, linear functions that are associated with each stimulus value during learning. When a new stimulus is evaluated, functions are activated based on the similarity between the new and associated stimuli. Then one rule is probabilistically selected and used to determine the response. Thus, similar to CX-COM, POLE involves a competitive selection mechanism. Consequently, an adaption of the competitive retrieval of exemplars as sketched in CX-COM is, in principle, able to explain some multimodal results POLE accounts for by assuming that sometimes an exception is recalled and adjusted according to a linear function. However, POLE stores different rules and assumes competition between these rules instead of exemplars, predicting that people also extrapolate in opposite directions depending on the exemplar they recall. CX-COM is unable to predict different extrapolation patterns on the same cue. Accordingly, although both models can predict multimodal response distributions, CX-COM and POLE differ on the items for which they predict a large variability. In CX-COM, variability is caused by training items that are activated by the same probe but differ in their criterion values; in POLE variability is caused by different functions associated with different parts of the stimulus space.

***Anchoring and Adjustment.*** On a more general level, CX-COM is also related to the idea that quantitative judgments are based on an anchoring and adjustment process (Tversky & Kahneman, 1974). According to the anchor and adjustment heuristic, people start making a judgment or an estimation by generating an initial value, the anchor. During the estimation process they then question whether the anchor provides an adequate judgment value and adjust their judgment until they are satisfied. Anchors can be internally generated values based on a memory processes or values that are externally provided in the environment (Chapman & Johnson, 2002; Epley & Gilovich, 2001;

Mussweiler & Strack, 2000). Similar to internally generated anchors, CX-COM assumes that a single exemplar is retrieved from memory and the value associated with this exemplar is adjusted based on the differences in cue values and beliefs about the relation between cues and the criterion. However, CX-COM does not allow for external anchors and their influence on the judgment process. In addition, the cue-based adjustment is based on the deviation between the features of the probe and the recalled exemplar and participants' assumptions about how these features relate to the criterion but not by further knowledge.

## Limitation and Future Work

We found that CX-COM accounted for judgments much better than the competing models in two experiments. However, in both experiments the number of exemplars in training was quite small and we ensured that participants memorized them very well. The open question remains of whether CX-COM still captures human judgments well if people retain more, but not necessarily intact exemplars in memory. One way people might react to more noisy exemplar representations is by giving more weight to the cue-abstraction component. Alternatively, it is possible that people will abstract prototypes or summary representations of exemplars that are clustered together. Then people might retrieve these prototypes instead of a single exemplar (Love, Medin, & Gureckis, 2004; Vanpaemel & Storms, 2008).

In both experiments we instructed participants to use the items they had learned during training to judge novel items during test (Olsson et al., 2006). We used these instructions because we aimed to provide a strong test of the retrieval assumptions underlying exemplar memory, that is, whether exemplar retrieval is integrative or competitive, and its interaction with cue knowledge. These instructions may limit the generality of CX-COM as a judgment model. However, the importance of exemplar-based processes in multiple-cue judgments studies without strategy instructions has been frequently demonstrated (Bröder & Gräf, 2018; Hoffmann et al., 2014, 2016; Juslin et al., 2008; Karlsson et al., 2007; McDaniel et al., 2018; Stillesjö, Nyberg, & Wirebring, 2019). Furthermore, a thorough, qualitative analysis of previous empirical evidence demonstrates that CX-COM can cover a broader variety of empirical findings (e.g. Juslin et al., 2008), including results that have been taken as evidence for transitions between exemplar memory and cue knowledge.

Although several memory models assume competitive retrieval, how many exemplars are recalled and combined before a response is given differs (e.g. Giguère & Love, 2013; Raaijmakers & Shiffrin, 1980). For example, the SAM (Search of Associative Memory) model assumes that memory images resulting from a competitive retrieval process are only partly restored. To restore a full memory image, several retrievals are necessary, implying that recalled images may be a combination of values from different memory items. In this work we tested the foundation of this idea: a model that considers only

one exemplar and a model that considers all exemplars to determine a response. If a subset of memory items is recalled, two scenarios are possible. If only exemplars with similar values are recalled from memory, the responses correspond to CX-COM's predictions. If very different values are recalled, they would be combined into one response similar to exemplar models with integrative retrieval. The focus on a competitive memory process in CX-COM allowed us to explore more complex and realistic memory processes in judgment research.

The present research has important implications for predictive models in the domain of quantitative judgments and evaluations. Although an integrative and a competitive retrieval mechanism as part of an exemplar model predict the same mean judgment values, the variance and actual shape of the distribution of response values might differ tremendously. Depending on the exemplars in memory that are activated in a specific context, a mean judgment value as predicted by integrative exemplar models might only be observed with a very low probability. Accordingly, the predictive power of classic exemplar models might be very low.

## Conclusions

We presented a new theory and cognitive model for quantitative judgments. CX-COM models how memories about specific exemplars and general beliefs about the relation of cues with criteria are integrated into a single judgment response. Most notably, CX-COM predicts multimodal response distributions and variability in judgments based on previously encountered exemplars and the similarity of these exemplars to the item under evaluation—an aspect in judgment behavior that has been largely neglected in research. In a quantitative model comparison CX-COM consistently outperformed all competitor models. In sum, CX-COM is a promising new model of the cognitive processes underlying quantitative judgments that allows researchers to derive distinct predictions for judgment behavior in various judgment situations.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*(4), 629–647.

Ashby, F. G., Alfonso-Reese, L. A., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.

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

Bröder, A., & Gräf, M. (2018). Retrieval from memory and cue complexity both trigger exemplar-based processes in judgment. *Journal of Cognitive Psychology*, *30*(4), 406–417.

Bröder, A., Gräf, M., & Kieslich, P. J. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making*, *12*(5), 491.

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of "rules". *Journal of Experimental Psychology: General*, *135*(2), 133.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108–132.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and biases: The psychology of intuitive judgment*, 120–138.

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

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.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Garner, W. R. (2014). *The processing of information and structure*. Psychology Press.

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, *110*(19), 7613–7618.

Hahn, U., & Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, *65*(2), 197–230.

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, *114*(1), 1–18.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, 70–84.

Herzog, S. M., & von Helversen, B. (2018). Strategy selection versus strategy blending: A predictive perspective on single-and multi-strategy accounts in multiple-cue estimation. *Journal of Behavioral Decision Making*, *31*(2), 233–249.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

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), 2242–2267.

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.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar

memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 924.

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

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607.

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

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

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PloS one*, *8*(12), e83528.

Lewandowsky, S., Murdock, B. B., et al. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25–57.

Lewandowsky, S., Roberts, L., & Yang, L.-X. (2006, Dec 01). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition*, *34*(8), 1676–1688. Retrieved from `https://doi.org/10.3758/BF03195930` doi: 10.3758/BF03195930

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), 309.

Macrae, C., Bodenhausen, G. V., Milne, A. B., Castelli, L., Schloerscheidt, A. M., & Greco, S. (1998). On activating exemplars. *Journal of Experimental Social Psychology*, *34*(4), 330 - 354.

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

McDaniel, M. A., Cahill, M. J., Frey, R. F., Rauch, M., Doele, J., Ruvolo, D., & Daschbach, M. M. (2018). Individual differences in learning exemplars versus abstracting rules: Associations with exam performance in college science. *Journal of Applied Research in Memory and Cognition*, *7*(2), 241–251.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of personality and social psychology*, *78*(6), 1038.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. *Choice, Decision, and Measurement:*

*Essays in Honor of R. Duncan Luce*, 347–365.

Nosofsky, R. M. (2014). The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization*, 18–39.

Nosofsky, R. M., & Palmeri, T. J. (1997). An Exemplar-based Random Walk Model of Speeded Classification. *Psychological Review*, *104*(2), 266.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53.

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.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 324.

Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences*, *8*(8), 378–386.

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. *Psychonomic Bulletin & Review*, *19*(4), 654–661.

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., & Wallsten, T. S. (2009). Random error in judgment: The contribution of encoding and retrieval processes. *Journal of Memory and Language*, *60*(1), 165–179. doi: 10.1016/j.jml.2008.08.003

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raaijmakers, J. G., & Shiffrin, R. M. (1980). Sam: A theory of probabilistic search of associative memory. *Psychology of Learning and Motivation*, *14*, 207–262.

Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual*

*Review of Psychology*, *43*(1), 205–234.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar-and rule-based theories of categorization. *Current Directions in Psychological Science*, *15*(1), 9–13.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley and Sons.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, *1*(1), 54–87.

Stillesjö, S., Nyberg, L., & Wirebring, L. K. (2019). Building memory representations for exemplar-based judgment: a role for ventral precuneus. *Frontiers in human neuroscience*, *13*.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, 201519157.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*(4), 732–749.

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental psychology*, *61*(1), 12.

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.

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

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*, 867.

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

## Appendix A: Similarities

In Experiment 1 participants rated how similar every training item was to every test item. Table A1 shows the aggregated results. The aggregated similarity ratings corresponded to the aggregated similarities predicted by the CX-COM model for all participants, ($r = 0.89, p < 0.01$). An analysis on the individual level confirms the results on the aggregate level (mean $r = 0.70, p < 0.01$) with every individual correlation being significant (individual $p$ values were corrected for multiple comparisons using the Bonferroni correction method).

Table A1

*Similarity ratings in Experiment 1*

| | Training item | | | | | |
|---|---|---|---|---|---|---|
| Judgment | 4 | 6 | 8 | 12 | 18 | 24 |
| test items | 3.1.2 | 2.1.2 | 3.2.2 | 1.1.3 | 2.2.3 | 1.2.3 |
| 3.1.1 | **7.00**(2.68) | 4.52(2.71) | 6.14(2.20) | 2.71(2.70) | 2.19(1.78) | 1.43(1.25) |
| 4.1.2 | **8.19**(1.47) | 4.67(2.82) | 5.67(2.76) | 2.05(1.75) | 2.14(1.53) | 1.33(1.56) |
| 2.1.1 | 4.76(2.64) | **6.90**(2.36) | 4.76(2.36) | 2.33(1.85) | 2.05(1.83) | 1.24(1.45) |
| 3.2.1 | 5.67(2.87) | 3.38(2.71) | **7.67**(1.93) | 1.62(1.69) | 2.67(1.80) | 2.48(2.20) |
| 4.2.2 | 6.38(2.75) | 3.76(2.41) | **7.95**(1.53) | 1.62(1.40) | 2.90(2.10) | 1.90(1.79) |
| 1.1.4 | 2.29(1.68) | 2.29(1.85) | 1.67(1.49) | **7.86**(1.82) | 5.29(1.95) | 5.29(2.24) |
| 2.2.4 | 1.90(1.84) | 2.62(2.25) | 2.67(2.13) | 5.33(2.33) | **8.14**(1.31) | 6.05(2.48) |
| 2.3.3 | 2.10(2.07) | 2.90(2.23) | 4.14(2.50) | 3.71(2.59) | **7.57**(1.29) | 5.14(2.92) |
| 1.2.4 | 1.71(2.03) | 1.76(1.61) | 2.43(1.43) | 6.29(1.65) | 6.67(1.71) | **8.24**(1.51) |
| 1.3.3 | 1.76(1.45) | 1.86(1.62) | 2.67(1.91) | 4.52(1.86) | 5.86(2.26) | **7.24**(2.23) |
| 2.3.2 | 2.81(2.14) | 5.05(2.65) | 4.33(2.54) | 2.10(2.26) | 5.00(2.26) | 2.76(2.41) |
| 2.1.3 | 5.38(2.69) | 6.62(2.87) | 4.00(2.59) | 6.71(2.45) | 6.95(2.01) | 4.62(2.50) |
| 1.3.2 | 3.62(2.25) | 2.67(1.77) | 3.57(2.25) | 3.00(1.70) | 3.95(2.16) | 5.43(2.56) |
| 2.3.2 | 1.76(1.41) | 2.76(1.79) | 3.67(2.13) | 1.43(1.91) | 3.76(2.07) | 2.76(2.30) |

*Note:* Participants' mean similarity ratings (and standard deviations) in Experiment 1. Highest perceived similarity is marked in bold.

## Appendix B: Fitting and Implementation Details

All analyses are done with the R programming language (R Core Team, 2015).

Table B1

*Model fit parameters*

| Parameter | CAM | Exemplar | CX-COM | RulEx-J | Baseline |
|---|---|---|---|---|---|
| Experiment 1 | | | | | |
| $w_1 = b_1$ (Dimension weight) | -.89 | 3.12 | .12 | 35.95 | - |
| $w_2 = b_2$ (Dimension weight) | 3.79 | 1.94 | .67 | -3.99 | - |
| $w_3 = b_3$ (Dimension weight) | 2.89 | 6.35 | .79 | -30.97 | - |
| $k$ (Intercept) | 1.79 | - | - | 1468.69 | - |
| $c$ (Sensitivity) | - | 11.42 | 1.58 | 1 | - |
| $\alpha$ (Cue-based adjustment) | - | - | -101.39* | - | - |
| $\beta$ (Model selection probability) | - | - | - | .5 | - |
| $\sigma^2$ (Error variance) | 5.45 | 6.51 | 2.93 | 5.33 | 8.51 |
| Mean in baseline model | - | - | - | - | 14.58 |
| Mean Deviance | 1,294 | 1,375 | 1,268 | 1,285 | 1,490 |
| Mean BIC | 1,321 | 1,397 | 1,294 | 1,317 | 1,501 |
| Number of parameters | 5 | 4 | 5 | 6 | 2 |
| Number of best fitted participants | 7 | 0 | 18 | 4 | 0 |
| Mean Deviance (CV fit) | 645 | 687 | 679 | 643 | - |
| Mean Deviance (CV prediction) | 654 | 691 | 647 | 650 | - |
| Number of best predicted participants (CV) | 2 | 0 | 14 | 13 | - |
| Experiment 2 | | | | | |
| $w_1 = b_1$ (Dimension weight) | 6.71 | .98 | .99 | 4.16 | - |
| $w_2 = b_2$ (Dimension weight) | 2.43 | .45 | .27 | 1.5 | - |
| $w_3 = b_3$ (Dimension weight) | 2.56 | .92 | .43 | 1.89 | - |
| $w_4 = b_4$ (Dimension weight) | 1.58 | .42 | .42 | .96 | - |
| $k$ (Intercept) | -8.7 | - | - | 66.97 | - |
| $c$ (Sensitivity) | - | 2.76 | 2.11 | 8.5 | - |
| $\alpha$ (Cue-based adjustment) | - | - | 33.64* | - | - |
| $\beta$ (Model selection probability) | - | - | - | .6 | - |
| $\sigma^2$ (Error variance) | 3.80 | 4.2 | 2.02 | 3.78 | 8.06 |
| Mean in baseline model | - | - | - | - | 17.26 |
| Mean Deviance | 1,077 | 1,119 | 1,033 | 1,075 | 1,400 |
| Mean BIC | 1,109 | 1,146 | 1,065 | 1,112 | 1,411 |
| Number of parameters | 6 | 5 | 6 | 7 | 2 |
| Number of best fitted participants | 9 | 2 | 19 | 1 | 0 |
| Mean Deviance (CV fit) | 535 | 558 | 518 | 536 | - |
| Mean Deviance (CV prediction) | 549 | 565 | 527 | 546 | - |
| Number of best predicted participants (CV) | 1 | 0 | 17 | 11 | - |

*Note:* Mean parameter values, Bayesian information criterion (BIC), and model descriptions. BIC and number of participants of the best fitting model are marked in bold.

*The high value for $\alpha$ in both experiments stems from a small number of participants with $\alpha$ values above 100. These participants were poorly fit by CX-COM and were not included in participants best fit by the model. The median for $\alpha$ is 2.86 in experiment 1 and 3.17 in experiment 2. The mean over participants best fit by the CX-COM model is 1.05 in experiment 1 and 6.92 in experiment 2.

The mean deviance of the cross validation (CV) are averaged across the two cross-validation sets (see Appendix B) for fits (CV fits) and predictions (CV predictions), number of best-predicted participants according to cross validation (CV).

For each model–participant combination we searched for the best fitting parameter setting by minimizing the models' negative log-likelihood. To find the best fitting model for every participant we used the Bayesian Information Criterion (BIC; Schwarz et al., 1978) to penalize more complex models.

All exemplar models (or model components) were fit to the data with the city-block distance (see Equation 11). Both exemplar and cue-abstraction models contain parameters reflecting the importance/attention given to the cue dimensions. To be able to fit both processes simultaneously we estimated only one set of dimension weights for both the exemplar retrieval process and the cue-based adjustment, so $w_i = b_i$ for each cue-dimension $i$. To do this, we freely estimated one weight for each cue dimension. In the cue-abstraction component, the adjustment was calculated according to the estimated weights. In the exemplar components we determined the attention weights $w_i$ and the sensitivity parameter $c$ by setting it to the sum of the absolute weights. We then calculated the attention weights by dividing the absolute weights of the respective dimensions by $c$. Hence, the attention weights in the exemplar process varied between 0 and 1 and summed up to 1 following the constraints usually assumed in exemplar models. Best fitting parameter values for both experiments are shown in Table B1.

For parameter estimation we used a combination of grid search and non-linear optimization. The grids had a step size of one and the overall size of the grid was informed by the true parameters of the functions underlying stimulus creation. In experiment 1 the borders of the grid were set to -10 and 10 and in experiment 2 to -20 and 15. These choices yielded 21 optimization searches for each model in experiment 1 and 46 in experiment 2. For each search, the starting parameter values were set to a random value between two subsequent grid values. Starting values outside the range of possible parameter values were ignored and set to the respective borders of the range instead. As optimization algorithm we used the "nlminb" function in the "stats" package of the R programming language (R Core Team, 2015).

***Cross Validation.*** Although CX-COM possesses the same number of parameters as the CAM and only one parameter more than the exemplar model, CX-COM may be more prone to overfitting than the CAM or the exemplar model. CX-COM is a mixture model which functionally traverses between pure exemplar and pure cue-abstraction predictions. Thus, similar to the mixture model RulEx-J Bröder et al. (2017), its functional complexity likely exceeds the functional complexity of pure exemplar and cue-abstraction models. To better understand whether the functional complexity is warranted given the data we conducted a cross validation.

For each participant, we split the set of observations for each test item into half and randomly assigned one half to the training set and the other half to the validation set in cross-validation. In Experiment 1, this resulted in splitting the observations into one set with 7 observations and one set with 8 observations. In Experiment 2, this resulted in two sets with 5 observations each. We then

estimated the model parameters' for each model and participant using participants' responses to items in the training set and predicted the responses from the validation set (and vice versa). Reported results are the mean of these two predictions.

## Appendix C: Beanplot for all items tested in Experiment 2

Figure C1 shows a beanplot for all items in Experiment 2.

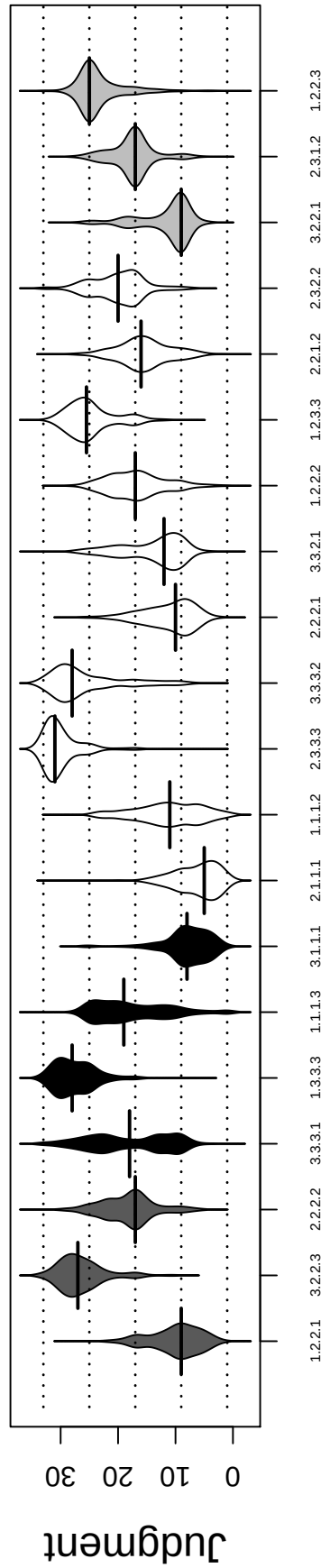


*Figure C1.* Participants' response distributions (as shown in Figure 7) for all items tested in Experiment 2. Dark gray distributions correspond to items with three most similar training items, black distributions to test items with two most similar training items, white distributions to item with one most similar training item, and light gray distributions correspond to training items repeated during the test phase. Thick dotted lines correspond to the criterion values of the training items; thin lines show the median of the distribution. The standard deviation of the kernel estimation was set to 1.06 (Scott, 1992).

## Appendix D: Post-hoc power analysis of multimodality within participant and item in Experiment 2

Within participants we only have 10 observations per item in experiment 2 and lacked the power to detect multimodality. Out of the 620 statistical tests on the participant/item level approximately 20% were significant ($p < .05$, not corrected). To better understand how many observations would have been needed to have enough power, we performed a post-hoc power analysis. Thereby we utilized the fact that CX-COM predicts a multimodal response distribution. We drew 10, 20, 50, 100, and 1000 samples (with 100 repetitions each) from the response distributions predicted by CX-COM for each participant/item combination. In case of 10 samples, multimodality was only detected in 17% of all tests, followed by 21%, 53%, 82%, and 100% in case of 1000 samples.

To test how often false positive results occur, we checked how likely a normal distribution is falsely identified as being multimodal by the dip test: Drawing 10 samples (same number as observations per participant and item in the experiment) with 1000 repetitions from normal distributions with variances of 1, 2, and 5 there were less than 0.2% false positive results in all three cases.

## Appendix E: Reanalysis of previous data and its limitations

We reanalyzed data from Hoffmann et al. (2014, 2016) who systematically investigated judgment strategies across different environments without strategy instructions. For all the different environments we fitted the CX-COM model, the CAM, and the exemplar model. In the environments from the 2016 paper CX-COM explains most participants best in the one-dimensional linear environment (22 out of 32 in the first and 24 out of 32 in the second variant) and in the multi-dimensional multiplicative environment (all 32), and it explains about half of the participants best in the multi-dimensional quadratic environment (16 out of 32). In the multi-dimensional linear environment the CAM is the best model with 18 out of 32 and CX-COM explains only 7 participants best. In the environments tested in the 2014 paper CX-COM is the best model in the multiplicative condition (267 out of 287) and the CAM is the best model in the linear condition (176 out of 287).

Unfortunately, these environments were not designed for tearing apart CX-COM from other models of human judgment. A model recovery suggested that CX-COM and the CAM could not be distinguished because CX-COM can make similar predictions as a cue abstraction model. Importantly, CX-COM and the CAM often only differ when predicting full response distributions instead of average responses. Therefore, it is necessary to observe many responses on the same test items to successfully recover CX-COM and contrast it with a CAM. Previous studies on judgment research, however, usually tested judgments for many test items but did not assess full response distributions for single items.

This data structure thus poses a problem for evaluating CX-COM's performance using previously published data. In the multi-dimensional linear environment from the 2016 paper, for example, CX-COM's recovery rate is around 50% while CAM's is around 90%. In contrast, in the multi-dimensional multiplicative condition the CAM can only be recovered in less than 40% while CX-COM can be recovered in more than 90% of all cases.