Justifying the judgment process affects neither judgment accuracy, nor strategy use

Janina A. Hoffmann*[†] Wolfgang Gaissmaier* [‡] Bettina von Helversen [†] [§]

Abstract

Decision quality is often evaluated based on whether decision makers can adequately explain the decision process. Accountability often improves judgment quality because decision makers weigh and integrate information more thoroughly, but it could also hurt judgment processes by disrupting retrieval of previously encountered cases. We investigated to what degree process accountability motivates decision makers to shift from retrieval of past exemplars to rule-based integration processes. This shift may hinder accurate judgments in retrieval-based configural judgment tasks (Experiment 1) but may improve accuracy in elemental judgment tasks requiring weighing and integrating information (Experiment 2). In randomly selected trials, participants had to justify their judgments. Process accountability neither changed how accurately people made a judgment, nor the judgment strategies. Justifying the judgment process only decreased confidence in trials involving a justification. Overall, these results imply that process accountability may affect judgment quality less than expected.

Keywords: judgment, accountability, cognitive processes.

1 Introduction

Providing a satisfying explanation for one's judgment plays a major role in professional life. Court decisions usually state the reasons for judgment, university teachers have to provide arguments for their grades upon request, and business decisions are evaluated by law by the degree they were taken on an informed basis, in good faith, and in the best interest of the company. Psychological research generally defines accountability as "the implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others" (Lerner & Tetlock, 1999, p. 255).Usually, two types of accountability are differentiated: outcome and process accountability (Langhe, Van Osselaer, & Wierenga, 2011; Siegel-Jacobs & Yates, 1996). Whereas performance evaluation on the basis of outcome usually seems to produce negative side-effects, performance evaluation based upon the judgment process can benefit performance in a range of tasks (Lerner & Tetlock, 1999; DeCaro, Thomas, Albert, & Beilock, 2011).

Yet, process accountability may prove advantageous only if people have to weigh and integrate all pieces of information. Process-accountable participants use more information to make a judgment (Kahn & Baron, 1995), but also consider irrelevant information more often (Siegel-Jacobs & Yates, 1996). Similarly, Langhe et al. (2011) found that process accountability improved performance only in multiple-cue judgment tasks in which people deliberately weigh and integrate information in a rule-based fashion, but not in tasks solved by retrieving past instances from memory.

Taken together, tasks demanding weighing and integrating information benefit from process accountability, but it remains unclear why tasks demanding memory retrieval do not. The current paper aims to understand when and why process accountability helps or hurts judgments. Specifically, holding people accountable for the judgment process may evoke a preference for thoroughly weighing and integrating information. We suggest that this strategy shift, in turn, influences how accurately the judgment task will be solved.

2 Judgment strategies in multiple-cue judgment tasks

In multiple-cue judgment tasks, the judge evaluates an object on a continuous scale using a number of attributes (or cues). When judging students' essays, for instance, the teacher determines the grade (the criterion) based on indicators of quality (i.e., the cues) such as coherent reasoning or good writing style.

Evidence has accumulated that people employ two kinds of judgment strategies: cue abstraction and exemplar memory (Juslin, Karlsson, & Olsson, 2008; Hoffmann, von Hel-

This research was supported by Swiss National Science Foundation Grant 100014_146169/1.

All material and the data can be downloaded from the Open Science Framework (Hoffmann, Gaissmaier, & von Helversen, 2017).

*Department of Psychology, Postbox 146, University of Konstanz, Universitaetsstrasse 10, 78 468 Konstanz, Germany. Email: janina.hoffmann@uni-konstanz.de.

[†] University of Basel

[‡] Max Planck Institute for Human Development

[§] University of Zurich

versen, & Rieskamp, 2014; von Helversen & Rieskamp, 2008). Cue abstraction strategies assume that people try to understand how each cue relates to the criterion, weigh each cue by its importance and then integrate them to a final judgment. For instance, teachers may emphasize the coherence of the reasoning over formal criteria. In contrast, exemplar-based strategies assume that people retrieve information about previously stored exemplars when judging a new instance. The higher the similarity of a stored exemplar to the to-be judged object, the more this past exemplar influences the final judgment. For instance, tutors could judge students' essays based on example cases they received from the professor.

Past research suggests that people select among those two strategies depending on task properties and the cognitive abilities of the decision maker (Hoffmann, von Helversen, & Rieskamp, 2013; Hoffmann et al., 2014; Juslin et al., 2008; Mata, von Helversen, Karlsson, & Cüpper, 2012; von Helversen, Mata, & Olsson, 2010). Specifically, people tend to rely on cue abstraction strategies in elemental judgment tasks in which the criterion is a linear function of the cues (Juslin et al., 2008; Hoffmann, von Helversen, & Rieskamp, 2016), and to rely on exemplar memory in configural judgment tasks in which the criterion is a non-linear function of the cues (Juslin et al., 2008; Hoffmann et al., 2013). Furthermore, putting a cognitive load on the decision maker limits the ease with which rules can be tested and motivates exemplar retrieval (Hoffmann et al., 2013) suggesting that participants processing the information more thoroughly may likewise engage in a qualitatively different judgment strategy.

3 Effects of process accountability on judgment strategies

How should process accountability interact with judgment strategies? Langhe et al. (2011) argued that process accountability boosts cue abstraction, but leaves exemplar memory unaffected. Specifically, process accountability may increase the motivation to thoroughly understand the decision process (De Dreu, Beersma, Stroebe, & Euwema, 2006; Langhe et al., 2011). Processing the available information more systematically may in turn benefit performance in tasks in which cue abstraction is a viable strategy. In line with this hypothesis, Langhe et al. (2011) found that process accountability increased judgment accuracy in an elemental judgment task. The consistency of the cue abstraction strategy in describing judgments explained this performance increase, given that participants scoring low on a rationality scale applied the cue abstraction strategy more consistently when they were process accountable.

It remains unclear, however, how process accountability affects exemplar memory. Langhe et al. (2011) reasoned that, if exemplars are automatically stored and retrieved from memory, processing the available information more systematically may not help. Consistent with this idea, judgment accuracy did not vary between process and outcome accountability in a configural, quadratic task (Langhe et al., 2011).

Alternatively, process accountability may induce a shift towards a cue abstraction strategy in both elemental and configural tasks. In line with this idea, awareness of the judgment process has been shown to foster a preference for rule-based processes in categorization (DeCaro et al., 2011; McCoy, Hutchinson, Hawthorne, Cosley, & Ell, 2014). Specifically, videotaping participants' performance hurts category learning in information-integration tasks, but not in rule-based tasks (DeCaro et al., 2011). Performance likely decreased because participants abandoned implicit strategies more often and considered two- and three-dimensional rules instead. Unfortunately, Langhe et al. (2011) did not investigate which judgment strategies underlie accuracy in the configural, quadratic task and current research still debates if people solve this task by storing exemplars or drop back to an unsuccessful cue abstraction strategy (Hoffmann et al., 2016; Olsson, Enkvist, & Juslin, 2006; Pachur & Olsson, 2012). As it stands, it is still an open question whether process accountability left exemplar memory unaffected or whether it motivated a higher reliance on cue abstraction.

4 Rationale of the experiments

The current experiments tested whether holding decision makers accountable for the judgment process counteracts exemplar-based processing and instead fosters cue abstraction. If so, process-accountable participants should approach a configural judgment task by cue abstraction and thus solve configural tasks less accurately (Experiment 1). In elemental tasks (Experiment 2), however, this preference for cue abstraction should help process-accountable participants solve the judgment task more accurately. To foreshadow our results, our experiments do not provide any support for the hypothesis that process accountability invokes a higher reliance on cue abstraction, neither in a configural exemplar-based task, nor in an elemental task. Justifications neither harmed judgments in a configural task nor benefitted judgments in an elemental task.

5 Experiment 1: Accountability in a configural judgment task

To test our prediction, we manipulated the need to justify one's judgment process while participants learned to solve a multiple-cue judgment task. In the accountability condition participants had to explain their judgments after randomly selected trials so that another person would be able to reproduce their judgments. Prompting justifications at random (rather than a single justification at the end) should motivate participants more in each single trial to explicitly reason about the judgment process. Further, providing a justification directly after the judgment reduces retrospection and increases validity of the justification (Lagnado, Newell, Kahan, & Shanks, 2006).

Like Langhe et al. (2011) we chose a configural task, but selected a multiplicative task that more reliably induces exemplar-based processes (Hoffmann et al., 2014, 2016). To pin down the strategy changes unique to justification, we compared the justification condition to one control condition without any accountability instruction and one with verbalization instructions because a mere verbalization of judgment processes may interfere with non-verbal processes too (Schooler, 2002; Deshon, Chan, & Weissbein, 1995; Schooler & Engstler-Schooler, 1990). Finally, we asked for confidence ratings after every trial, although we did not specify a hypothesis in advance for these.

5.1 Method

5.1.1 Participants

Out of 153 participants from the participant pool of the Max-Planck-Institute for Human Development, Berlin, we had to discard 9 incomplete data sets due to error, leaving a sample of 144 participants (80 female, $M_{\text{Age}} = 25.4$, $SD_{\text{Age}} = 3.3$). Participants received an hourly wage of 13 € as well as a performance-dependent bonus ($M = 2.90$ €, $SD = 0.84$ €).

5.1.2 Design and Material

In the adapted judgment task from Hoffmann et al. (2016), participants estimated the toxicity of a bug (the criterion) on a scale from 0 to 50 mg/l. The criterion y was predicted by four quantitative cues, x_1,..., x_4 with cue values ranging from 0 to 5, that were combined multiplicatively:

$$y = \frac{4x_1 + 3x_2 + 2x_3 + x_4 + 2x_1x_2x_3 + x_2x_3x_4}{8.5} \quad (1)$$

We used the same items in the judgment task as in previous studies (Hoffmann et al., 2014, 2016, Appendix A). The items were selected so that an exemplar strategy allowed to more accurately judge the old training items than the cue abstraction strategy and that the new validation items discriminated among the judgment strategies.

The pictorial stimuli displayed bugs varying on four visual features: the length of their legs, their antennae, and their wings, and the number of spots on their back. These visual features could be used to predict the bug's toxicity. The cues x_1,..., x_4 were randomly assigned to the visual features (e.g., antennae). Higher cue values were always associated with more salient visual features. For instance, a cue value of zero on the cue 'legs' corresponded to a bug without (visible) legs, whereas a bug with a cue value of five had long legs.

5.1.3 Procedure

Participants were first instructed that they will learn to predict the toxicity of different bugs during the training phase. Additionally, participants in the justification condition were informed that they would have to justify their judgments so that another person could make the same judgments based upon their descriptions (see Appendix B for verbatim translations of the instructions). In the verbalization condition, participants were informed that they will have to subdivide their judgments into its components.

Next, we introduced a practice task to help participants imagine what information they (or another person in the justification condition) would need to accurately judge the bugs' toxicity. In this task, participants saw a bug with different cues and had to indicate the information they would need to accurately judge the bugs' toxicity based only upon a verbal description.

The subsequent judgment task consisted of a training and a test phase. During training, participants learned to estimate the criterion values for 25 training items. In each trial, participants first saw a bug and estimated its toxicity. Next, participants rated their confidence by estimating how much their answer deviated from the correct judgment.[1] Finally, they received feedback about the correct value, their own estimate, and the points earned. Training ended after 10 training blocks, with 25 training items presented in random sequence in each block.

In 20 of these 250 judgment trials, the experimental trials, participants justifying their judgment had to explain their judgment so that another person could make the same judgment, but without mentioning the specific judgment value. Participants in the verbalization condition indicated how much each cue contributed to the total toxicity. Verbalizations and justifications occurred randomly twice in each training block, directly after the judgment (see Figure 1). Thus participants could not know beforehand in which trials they would need to justify (or verbalize) their judgment.

In the subsequent test phase, participants judged 15 new validation items four times and indicated their confidence but did not receive any feedback. Further, participants neither verbalized, nor justified their judgments.

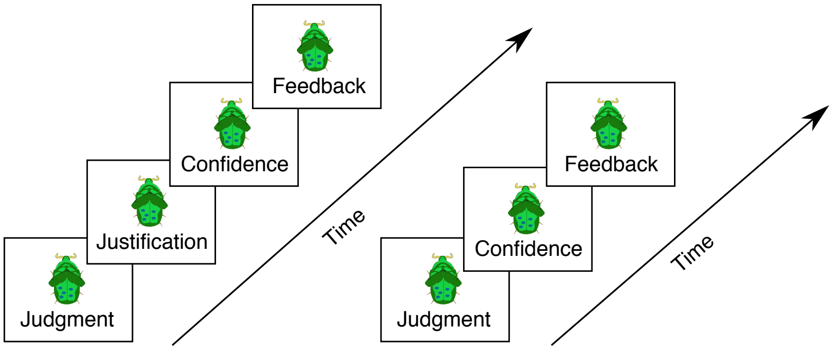
To motivate a high performance, participants could earn points in every trial. The points earned were a truncated quadratic function of the deviation of their judgment j from the criterion y:

$$\text{Points} = 20 - \frac{(j - y)^2}{7.625} \quad (2)$$

This incentivization scheme was communicated to participants in the instructions: "Every correct estimate will earn you 20 points. Almost correct estimates will earn you less

[1] We collected response and processing times, but we did not postulate any effect of justification on response times, nor did we analyze the response times.

Figure 1: Trial sequence for experimental (left sequence) and control trials (right sequence). In the experimental trials, participants in the justification condition had to justify their judgment after they made a judgment, whereas participants in the verbalization condition indicated how much each cue contributed to the total amount of toxicity.



points. If you deviate from the correct value by more than 12 points, you will not earn any points." At the end of the experiment, the points earned were converted to a monetary bonus (4000 points = 1 €). In addition, participants earned a bonus of 2 € if they reached 80% of the points in the last training block (corresponding to less than 5.5 RMSD [root mean square deviations]). Verbalization questions and justifications were incentivized, too. Participants in the verbalization condition could gain 20 additional points for each verbalization question if the importance assigned to each cue summed up to their judgment. Participants in the justification condition could win one Amazon voucher worth 50 €with higher chances of winning the more closely another person approximated the judgment based upon their justification.[2]

5.2 Results

Bayesian analyses were performed in R (R Core Team, 2016) to quantify evidence for and against the null hypothesis with Bayes Factors (BF, calculated with the BayesFactor Package and the specified defaults priors in this package Morey, Rouder, Jamil, & Morey, 2015). BFs express the relative likelihood of one hypothesis over another one in light of the data. BFs above 3 provide moderate evidence, BFs above 10 provide strong evidence for the alternative hypothesis (Lee & Wagenmakers, 2014; Jeffreys, 1961). BFs below 1 provide evidence for the null hypothesis.

5.2.1 Does justification decrease judgment performance?

Participants learned to solve the judgment task equally well in all conditions (see Figure 2 and Table 1 for descriptive statistics). Most participants reached the learning criterion, and learning success did not vary between conditions (BF = 0.058, using a Bayesian test for contingency tables assuming independent multinomial sampling with a Gunel and Dickey prior with prior concentration set to 1).

To test if justification decreased judgment accuracy compared to the verbalization and the control condition, we performed a repeated measures Bayesian ANOVA on judgment error, measured in RMSD between participants' judgments and the correct criterion in each block, with the factors training block and condition.[3] Judgment error dropped in all conditions from the first to the last training block ($BF_{\text{Block},0}$ > 10000), but justifying judgments did not increase judgment error ($BF_{\text{Cond},0}$ = 0.182), nor did the need to justify or verbalize judgments change learning speed ($BF_{\text{Block x Cond,Block}}$ = 0.031). A corresponding Bayesian ANOVA on judgment error in test found no evidence that justifying judgments decreased judgment accuracy more than in the control conditions ($BF_{\text{Cond},0}$ = 0.180). In sum, participants learned to make accurate judgments with more training blocks, but justifications did not decrease judgment accuracy in training or test.

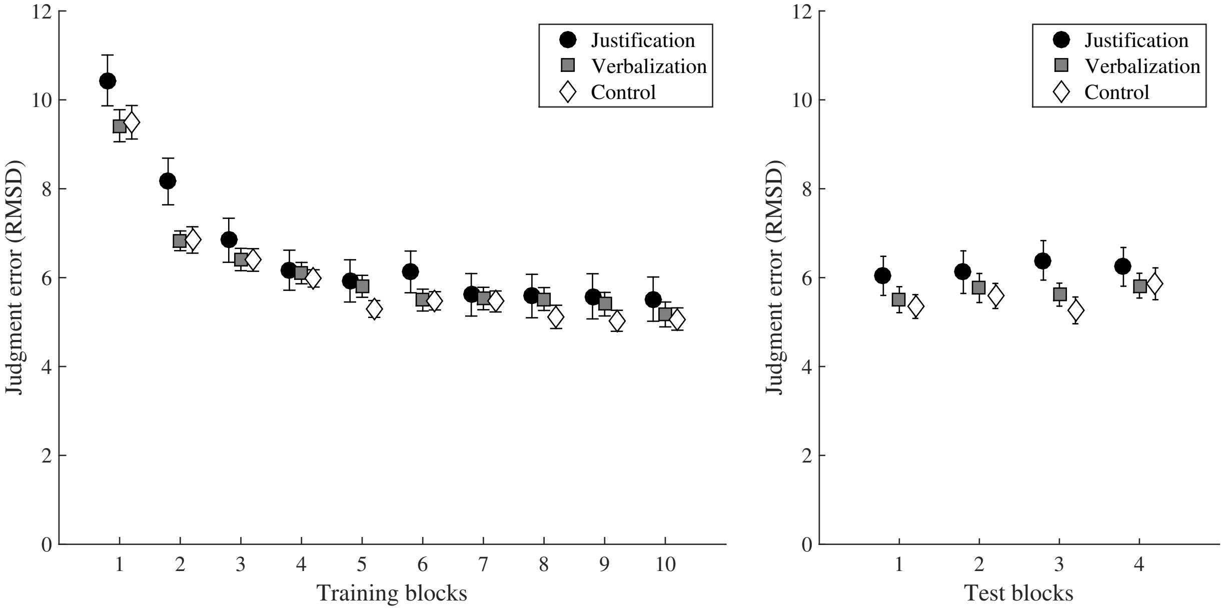
5.2.2 Judgment strategy and accuracy

To better understand on which judgment strategy participants based their judgment, we fitted three judgment models to participant's judgments in training and predicted their judgments in test (see Appendix C and Hoffmann et al., (2014, 2016)): a cue abstraction, an exemplar, and a guessing model. As expected, most participants were best described by an exemplar model in the control condition (see Table 2 for strategy classification and performance by strategy), but strategies did not change depending on the condition

[2] To measure how closely another person approximated the judgment of the participant, we randomly selected five justifications for each participant (320 justifications in total). In a later study, a rater judged the bug based upon the justification and the corresponding picture. All justifications were randomly interspersed, stated judgment values replaced by "XX" and the rater was aware that justifications were generated by different participants.

[3] In the Bayesian ANOVA, g-priors are assumed for the effects and independent scaled inverse-chi-square priors with one degree of freedom and a corresponding scaling parameter r are placed on g (Morey et al., 2015; Rouder, Morey, Speckman, & Province, 2012) with r = 1/2 for the fixed and r = 1 for the random effects.

FIGURE 2: Judgment error in the training phase (left plot) and the test phase (right plot) measured in Root Mean Square Deviations (RMSD) in Experiment 1, separately for participants in the justification (dark grey circles), the verbalization (light grey squares), and the control condition (white diamonds). Error bars show ± 1 *SE*.



(BF = 0.002, Bayesian test for contingency tables assuming independent multinomial sampling).

Did the chosen strategy influence how accurately and consistently participants judged the test items? To quantify how strategy choice affected accuracy in the test, we included judgment strategy as an independent variable in the ANOVA on accuracy. Participants best described by guessing were excluded in all analyses involving strategy choice. Overall, participants classified to the exemplar model were more accurate in the test ($\mathrm{BF}_{\mathrm{Strategy},0}$ = 2379), but justification did not affect judgment error ($\mathrm{BF}_{\mathrm{Cond},0}$ = 0.143). Finally, people assigned to a cue abstraction model in the justification or verbalization condition did not make more errors in the test than participants assigned to the cue abstraction model in the control group ($\mathrm{BF}_{\mathrm{Strategy\ x\ Cond,Strategy}}$ = 0.027). Because justification may also affect how consistently participants judge the same items (Siegel-Jacobs & Yates, 1996), we performed a corresponding analyses on judgment consistency, measured as the average correlation between the judgments in the test blocks. The results on consistency mimic the pattern for accuracy. Cue abstraction users made less consistent judgments in the test ($\mathrm{BF}_{\mathrm{Strategy},0}$ = 8.2), but neither justification ($\mathrm{BF}_{\mathrm{Cond},0}$ = 0.069) nor its interaction with judgment strategy affected consistency ($\mathrm{BF}_{\mathrm{Strategy\ x\ Cond,Strategy}}$ = 0.028). In sum, justification did not lead to a shift to cue abstraction and did not change how accurately or consistently participants judged the new items.

5.2.3 Post-hoc analyses of confidence ratings and justifications

So far, we found no evidence that justifying the judgment process alters judgment performance. Confidence ratings and the stated justifications may provide further information about whether our prompt to justify one's judgment changed the judgment process.

Previous research considering confidence ratings suggests that participants who have to justify their judgments are on average better calibrated (Siegel-Jacobs & Yates, 1996) and less overconfident (Tetlock & Kim, 1987). If justifications had any effect on judgment strategy in our experiment, the need to justify one's judgment should at minimum reduce confidence for judgments on the same trial because confidence ratings directly followed justifications. Justifications may also affect confidence in trials before or after the justification, but this would be a smaller effect. Hence, participants in the justification condition should have indicated that their judgments further deviated from the correct criterion in trials in which they had to justify their judgment than in preceding or subsequent trials without the justification prompt. To test this possibility by considering only relative decrements in judgment confidence, we first z-standardized participants' confidence ratings for each item across all participants and trials in the training phase. Next, we averaged these confidence ratings for each participant, separately for trials preceding the justification (or the verbalization ques-

Table 1: Performance in Experiment 1 (Configural Task) and Experiment 2 (Elemental Task). Standard Deviations in Parentheses.

	Experiment 1			Experiment 2	
	Justification (*n* = 49)	Verbalization (*n* = 47)	Control (*n* = 48)	Justification (*n* = 55)	Control (*n* = 55)
Training session					
Error first block	10.4 (4.0)	9.4 (2.5)	9.5 (2.6)	9.8 (3.0)	8.9 (2.3)
Error last block	5.5 (3.5)	5.2 (1.9)	5.1 (1.7)	5.3 (2.0)	5.0 (2.0)
Bonus *n*	39 (79.6%)	38 (80.9%)	42 (87.5%)	38 (69.1%)	44 (80.0%)
Guessing *n*	4	3	2	1	2
Test session					
Mean error	6.2 (3.0)	5.7 (1.8)	5.5 (1.9)	6.2 (1.9)	5.8 (1.7)
Mean confidence					
Training first block	4.5 (2.8)	3.8 (1.8)	3.8 (1.6)	6.1 (4.0)	5.3 (2.3)
Training last block	3.2 (1.5)	3.4 (1.8)	3.3 (1.2)	4.2 (2.1)	4.0 (1.6)
Test	3.3 (1.5)	3.5 (1.7)	3.4 (1.5)	4.4 (1.8)	4.6 (2.5)
z-Confidence					
Pre-trial	0.02 (0.71)	−0.01 (0.71)	−0.03 (0.55)	0.06 (0.68)	−0.09 (0.50)
Trial	0.21 (0.90)	−0.01 (0.68)	0.01 (0.55)	0.27 (0.85)	−0.09 (0.49)
Post-trial	0.01 (0.63)	−0.04 (0.69)	−0.04 (0.56)	0.01 (0.57)	−0.09 (0.49)

Note: Error in the judgment tasks was measured in root mean square deviation. The bonus reports the number (or percentage) of participants reaching the learning criterion. Confidence ratings asked participants how far their judgment deviated from the correct judgment.

tion), trials that contain a justification, and trials after the justification (we refer to the different trials as trial type). For participants in the control condition, we randomly selected two trials in each training block and used the trials preceding or following it as a comparison.

Table 1 depicts z-standardized confidence ratings in each condition, separately for each trial type. Descriptively, participants in the justification condition are on average less confident on trials in which they had to justify their judgment, than on trials in the control or the verbalization condition; still, the difference is small, as indicated by an increase of only 0.2 *SD* compared to the average confidence. A repeated measures ANOVA suggested that participants in the justification condition were not generally less confident about their judgments (including all trials, $\mathrm{BF}_{\mathrm{Cond},0}$ = 0.33). However, participants were less confident in trials with a justification than in trials preceding or following a justification ($\mathrm{BF}_{\mathrm{Trial\ type},0}$ = 22.5). How strongly confidence changed as a function of trial type depended on the condition ($\mathrm{BF}_{\mathrm{Cond\ x\ Trial\ type},0}$ = 9.8), but BFs could not clearly distinguish whether only trial type influences confidence or whether the trial type plays a stronger role in specific conditions ($\mathrm{BF}_{\mathrm{Cond\ x\ Trial\ type},\mathrm{Trial\ type}}$ = 0.435).

In a follow-up analysis, we therefore put equality constraints on trial type, separately for each condition, that is, we conducted the same repeated measures analysis but did not allow confidence to vary in one of the conditions. Next, we compared this constrained model against the unconstrained repeated measures analysis. BF above 1 indicate a preference for the constrained model. Assuming no change in confidence was acceptable for the control condition (BF = 5.2) and the verbalization condition (BF = 10.4), but not in the justification condition (BF < 0.001). In sum, this result suggests that participants who had to justify their judgment were less confident in trials including this justification, but participants who had to verbalize their judgment were not. (Of course, the control condition showed no such effect of trial type, by design, since trials were selected randomly.)

Finally, we rated how often participants provided reasons in their justifications (a binary rating) and which reasons they provided (see Appendix D for methodology and summary statistics). Providing reasons more often did not correlate with participants' success in solving the judgment task at the end of training (*M* = 70.2% of trials, *SD* = 38.1% of

TABLE 2: Performance and Strategy Consistency Separately for Participants Classified to Each Strategy (Cue Abstraction or Exemplar) in Experiment 1 (Configural Task) and Experiment 2 (Elemental Task). Standard Deviations in Parentheses.

	Experiment 1			Experiment 2	
	Justification	Verbalization	Control	Justification	Control
Strategies					
Guessing	1 (2 %)	0 (0 %)	0 (0 %)	2 (3.6 %)	1 (1.8 %)
Cue abstraction	24 (49 %)	17 (36.2 %)	21 (43.8 %)	43 (78.2 %)	48 (87.3 %)
Exemplar	24 (49 %)	30 (63.8 %)	27 (56.2 %)	10 (18.2 %)	6 (10.9 %)
Test session (Mean Error)					
Cue abstraction	6.9 (3.8)	7.1 (1.7)	6.2 (2.4)	6.3 (1.9)	5.5 (1.6)
Exemplar	5.3 (1.5)	4.8 (1.2)	5.0 (1.2)	5.6 (1.1)	7.4 (2.1)
Consistency *r*					
Cue abstraction	0.87 (0.37)	0.81 (0.33)	0.85 (0.39)	0.83 (0.39)	0.84 (0.32)
Exemplar	0.88 (0.48)	0.90 (0.32)	0.89 (0.38)	0.81 (0.33)	0.74 (0.41)

Note: Error in test session was measured as the root mean square deviation.

trials, r = 0.112). In a linear model, we predicted judgment error during training with the percentage of reasons stated for participants in the justification condition. The linear model indicated only that participants were more accurate in later training blocks ($BF_{Block,0} > 10000$), but the percentage of reasons stated did not influence judgment error ($BF_{Block+Reason,Block} = 0.602$), nor its interaction with training blocks ($BF_{Block*Reason,Block+Reason} < 0.001$). Also, a quality index, expressing how much information participants provided in their justifications, did not predict judgment accuracy at the end of training (M = 0.43, SD = 0.19, r = -0.029).

5.3 Discussion

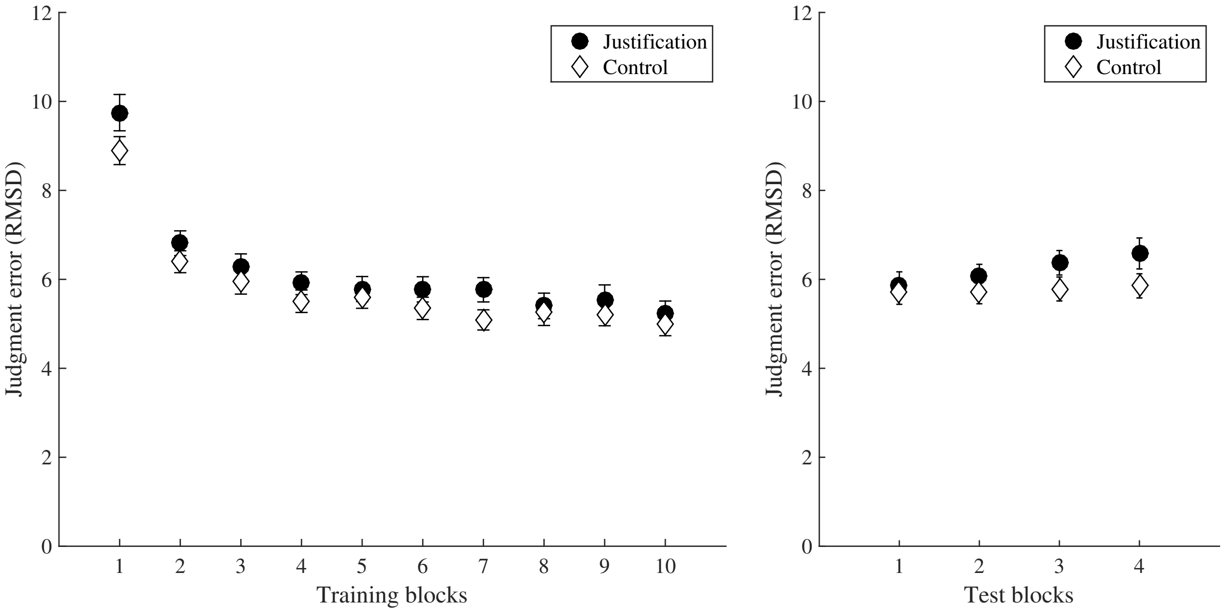
In sum, neither process accountability nor verbalization decreased judgment accuracy in the configural task compared to a control group receiving only outcome feedback. Furthermore, process accountable participants did not shift more towards cue abstraction strategies, contradicting our initial hypothesis and previous work in category learning (DeCaro et al., 2011). Potentially, this shift is more pronounced in categorization because participants can form explicit if-then rules based on one or more cues, whereas the cue abstraction strategies in judgment demand additive linear integration of cues. Exploratory analyses indicated that participants were slightly less confident after a justification, suggesting that participants at least reconsidered their judgment strategy. Our results resonate better with the finding that justifying the judgment process compared to justifying the outcome does not affect accuracy in configural, quadratic tasks (Langhe et al., 2011). In combination, these results hint at the interpretation that justifying one's judgment process does not interfere with more automatic retrieval from exemplar memory (Langhe et al., 2011).

If automatic retrieval of exemplars underlies the null effect of process accountability, one would expect process accountability to improve judgments in an elemental task that is better solved by cue abstraction. Yet, the beneficial effects of process accountability may be also overstated. In this vein, Siegel-Jacobs and Yates (1996) found that holding participants accountable for the process failed to affect judgment accuracy and only improved calibration (Exp. 1) or discrimination (Exp. 2). We address this question in experiment 2.

6 Experiment 2: Accountability in an elemental judgment task

In elemental judgment tasks, the benefits of process over outcome accountability are well documented (Langhe et al., 2011; Ashton, 1990, 1992). In three experiments, Langhe et al. (2011) provided convincing evidence that justifying the judgment process improves accuracy more than justifying the outcome. Similarly, stating reasons for one's judgment can promote a higher judgment accuracy even in the absence of social pressure (Ashton, 1990, 1992). Strategy preferences at the end of training unlikely account for this improvement because the majority of participants is best described by cue abstraction (Hoffmann et al., 2014). Still, process-accountable participants may develop a preference for cue abstraction earlier in training, as a consequence settle on their final judgment policy more quickly and apply the cue

Figure 3: Judgment error in the training phase (left plot) and the test phase (right plot) measured in Root Mean Square Deviations (RMSD) in Experiment 2, separately for participants in the justification (dark grey circles) and the control condition (white diamonds). Error bars show ± 1 *SE*.



abstraction strategy more consistently (Ashton, 1990, 1992; Langhe et al., 2011). In Experiment 2, we expected that process-accountable participants apply the cue abstraction strategy more consistently compared to a control condition without accountability and, hence, should make more accurate judgments in an elemental judgment task.

6.1 Method

6.1.1 Participants

A hundred-ten participants (58 females, M_{Age} = 25.6, SD_{Age} = 6.0) from the University of Basel received an hourly wage of 20 CHF (Swiss Francs) for their participation as well as a performance-dependent bonus (*M* = 5.49 CHF, *SD* = 1.59 CHF).

6.1.2 Material, Design, and Procedure

Compared to Experiment 1, we changed the function relating the cues to the criterion. Specifically, the judgment criterion y was a linear, additive combination of all cues:

$$y = 4x_1 + 3x_2 + 2x_3 + x_4 \tag{3}$$

The monetary incentive was converted to Swiss Francs (1500 points = 1 CHF) and participants earned additionally 3 CHF if they reached 80 % of the points in the last training block.

6.2 Results

6.2.1 Does justification increase judgment performance?

Participants on average learned to solve the judgment task well and justifications did not affect the number of participants reaching the learning criterion (BF = 0.475, see Table 1 for descriptive statistics and Figure 3). To investigate if justifying one's judgment improved judgment accuracy, we performed a repeated measures Bayesian ANOVA. This analyses suggested that, on average, judgment error dropped in both conditions from the first to the last training block ($\text{BF}_{\text{Block,0}}$ > 10000). Yet, BFs did not provide enough support for or against an undirected effect of justification ($\text{BF}_{\text{Block + Cond,Block}}$ = 0.560). Therefore, we tested more strictly the directional hypothesis by setting order constraints. This test rejected the idea that justification increases judgment accuracy (BF = 0.209). In addition, justifying one's judgment did not speed up learning compared to the control condition ($\text{BF}_{\text{Block x Cond,Block}}$ = 0.003). In the test, a directional Bayesian t-test also rejected the idea that justification enhanced judgment accuracy compared to the control group ($\text{BF}_{\text{Cond,0}}$ = 0.091). In sum, participants held accountable for the judgment process did not outperform participants in the control condition in training or in test.

6.2.2 Judgment strategy and accuracy

As expected, most participants were best described by a cue abstraction model (see Table 2), but no more participants were best described by cue abstraction in justification than in the control condition (BF = 0.056). In addition, process-accountable participants who followed a cue abstraction strategy did not make more accurate or more consistent judgments in the test. Including judgment strategy in the ANOVA on judgment error indicated neither that participants classified to cue abstraction were more accurate ($BF_{\text{Strategy},0}$ = 0.379), nor that justification improved judgment accuracy ($BF_{\text{Cond},0}$ = 0.393), nor that judgment strategy affected judgment accuracy differently depending on justification ($BF_{\text{Strategy x Cond},0}$ = 0.749). Similarly, how consistently participants judged the test items was not influenced by judgment strategy ($BF_{\text{Strategy},0}$ = 0.740), justification ($BF_{\text{Cond},0}$ = 0.205), or the interaction ($BF_{\text{Strategy x Cond},0}$ = 0.072). In sum, these results suggest that process accountable participants were no better described by a cue abstraction strategy, nor did the pursued strategy increase the accuracy and consistency of process accountable participants — potentially because the majority of participants was classified as using the cue abstraction strategy.

6.2.3 Post-hoc analyses of confidence ratings and justifications

Descriptively, justifying one's judgment reduced confidence directly after the justification; still, the effect was small. A repeated measure ANOVA did not indicate that justifications made participants less confident per se ($BF_{\text{Cond},0}$ = 1.2), but all participants were less confident directly after a justification ($BF_{\text{Trial},0}$ = 68.8). Importantly, participants justifying their judgment were less confident directly after the justification, but participants in the control condition were not ($BF_{\text{Trial Type x Cond,Trial Type}}$ = 140.9).

Analysis of the justifications indicated that success on the judgment task neither correlated with how often participants provided reasons (M = 56.2%, SD = 40.8%, r = 0.021), nor with the quality of the reasons stated (M = 0.38, SD = 0.23, r = −0.158). Furthermore, predicting judgment error across training in a linear model suggested that participants were more accurate only in later training blocks ($BF_{\text{Block},0}$ > 10000), but the percentage of reasons stated did not influence judgment error ($BF_{\text{Block+Reason,Block}}$ = 0.403), nor its interaction with training blocks ($BF_{\text{Block*Reason,Block+Reason}}$ < 0.001).

7 General Discussion

Giving reasons for decisions is a common duty in professional life. Such justifications may give insight into the judgment process and have been implemented as tools to improve judgment quality. Yet, our major results indicate that asking for a justification affects the decision process and judgment quality less than expected (Langhe et al., 2011). In two experiments, participants justified their judgments after randomly selected learning trials. In the first experiment, we expected justifications to encourage a higher reliance on cue abstraction and, consequently, harm performance in a configural judgment task that is better solved by exemplar memory. In a second experiment, we expected justifications to prove beneficial in an elemental task in which a cue abstraction strategy leads to a better performance. Yet, in both experiments, justifications did not encourage a more consistent use of a cue abstraction strategy, nor did justifications impede or profit judgment accuracy.

Our null results contradict the previously found beneficial effects of process accountability (Langhe et al., 2011; Lerner & Tetlock, 1999; but see Siegel-Jacobs & Yates, 1996). These previous studies mostly contrasted process with outcome accountability, whereas our study distinguished process accountability from a judgment process without accountability instructions. Matching our findings, a few previous studies found no evidence that process accountability benefits accuracy more than a no-accountability control in rule-based tasks (DeCaro et al., 2011; Siegel-Jacobs & Yates, 1996). Jointly considered, these results hint at the interpretation that outcome accountability worsens judgment performance and causes the difference between process and outcome accountability in elemental tasks.

Alternatively, the few justifications required may not have motivated participants enough to change their judgment policy compared to previous research (DeCaro et al., 2011). In both experiments, participants justifying their judgment process were not more likely to adopt a cue abstraction strategy, nor were their judgments more consistent. Yet, process accountable participants were slightly less confident directly after a justification, indicating that justifications only made people doubt their judgments. In addition, a lack of insight into one's own judgment policy may hinder a change towards cue abstraction (Haidt, 2001; Lerner & Tetlock, 1999; Nisbett & Wilson, 1977). Matching this idea, the quality of justifications did not correlate with judgment accuracy in our study and participants mentioned mostly superficial characteristics instead of deeply reflecting upon the judgment process. Potentially, asking more fine-grained questions about the judgment process may help participants to accurately reflect on, and ultimately change, their judgment policy (Lagnado et al., 2006).

Another limitation is potentially that we incentivized every trial and offered a bonus for reaching the learning criterion. First, some studies combined the possibility to win a bonus with social pressure to induce outcome accountability (DeCaro et al., 2011) and thus all our judgment tasks may involve some aspects of outcome accountability as well. Second, the chance to win a bonus itself (compared to losing a bonus) may induce a promotion focus and change how participants approach a judgment task (Grimm, Markman, Maddox, & Baldwin, 2008; Maddox, Baldwin, & Markman, 2006). In this vein, categorization research has found that participants who gain points on every trial and expect a bonus were closer to the optimal reward criterion than participants who expected to lose their bonus (Markman, Baldwin, & Maddox, 2005). Different incentivization schemes may alter how effectively people solve a judgment task, too, but research on incentivizations in judgment is rare (Ashton, 1990).

The impact of process accountability likely depends on its implementation, too. Past manipulations ranged from announcing a later report to a final interview to videotaping the judgment process (Langhe et al., 2011; DeCaro et al., 2011). Those manipulations vary in the frequency and timing of expected justifications or the social pressure involved. For instance, we induced social pressure by explaining that justifications will be reviewed by another person, but an expected interview with another person may have increased social pressure more strongly. Future research shall investigate more systematically which factors make people reliably feel accountable for the decision process and thereby aid practioners to successfully implement justifications as tools improving decision quality.

Taken together, our experiments provide little support for the common idea that providing a satisfying explanation towards others makes people weigh and integrate all information more systematically, which could improve or decrease performance depending on the structure of the decision task.

References

Ashton, R. H. (1990). Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research*, *828*, 148–180.

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes*, *52*(2), 292–306.

Busemeyer, J. R., & Wang, Y.-M. (2000, mar). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189.

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

De Dreu, C. K. W., Beersma, B., Stroebe, K., & Euwema, M. C. (2006). Motivated information processing, strategic choice, and the quality of negotiated agreement. *Journal of Personality and Social Psychology*, *90*(6), 927–943.

DeCaro, M. S., Thomas, R. D., Albert, N. B., & Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General*, *140*(3), 390–406.

Deshon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal Overshadowing Effects on Raven's Advanced Progressive Matrices: Evidence for Multidimensional Performance Determinants. *Intelligence*, *21*, 135–155.

Grimm, L. R., Markman, A. B., Maddox, W. T., & Baldwin, G. C. (2008). Differential effects of regulatory fit on category learning. *Journal of Experimental Social Psychology*, *44*, 920–927.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Hoffmann, J. A., Gaissmaier, W., & von Helversen, B. (2017). *Justification in Judgment.*

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242–2261.

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

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

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 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), 133–156.

Kahn, B. E., & Baron, J. (1995). An Exploratory Study of Choice Rules Favored for High-Stakes Decisions. *Journal of Consumer Psychology*, *4*(4), 305–328.

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006, may). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183.

Langhe, B. D., Van Osselaer, S. M. J., & Wierenga, B. (2011). The effects of process and outcome accountability on judgment process and performance. *Organizational Behavior and Human Decision Processes*, *115*(2), 238–252.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275.

Maddox, W. T., Baldwin, G. C., & Markman, A. B. (2006). A test of the regulatory fit hypothesis in perceptual classification learning. *Memory & Cognition*, *34*(7), 1377–1397.

Markman, A. B., Baldwin, G. C., & Maddox, W. T. (2005). The Interaction of Payoff Structure and Regulatory Focus in Classification. *Psychological Science*, *16*(11), 852–855.

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188–1201.

McCoy, S. K., Hutchinson, S., Hawthorne, L., Cosley, B. J., & Ell, S. W. (2014). Is pressure stressful? The impact of pressure on the stress response and category learning. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 769–81.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). *BayesFactor: Computation of Bayes Factors for Common Designs.* Retrieved from `https://cran.r-project.org/package=BayesFactor`

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, *84*(3), 231–259.

Nosofsky, R. M., & Zaki, S. R. (1998, jul). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*(4), 247–255.

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006, nov). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1371–1384. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/17087590http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.32.6.1371`

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

R Core Team. (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012, oct). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Schooler, J. W. (2002, dec). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*(8), 989–997.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990, jan). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71.

Siegel-Jacobs, K., & Yates, J. (1996). Effects of Procedural and Outcome Accountability on Judgment Quality. *Organizational Behavior and Human Decision Processes*, *65*(1), 1–17.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*(4), 700–709.

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), 867–889.

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

Appendix A: Items used in judgment task

Appendix B: Instructions

Below we list the instructions participants received in each condition. Instructions were translated into English in a verbatim fashion.[4]

Instructions in the justification condition

"In this task, it is of particular importance that you not only make a judgment, but are also able to well justify and explain these judgments. For this reason, we will randomly prompt you after some of your judgments to accurately justify and explain your judgment in written form so that another person is able to reproduce your judgment and reaches the same judgment. The other person will see the bug and your justification and likewise makes a judgment based upon this information. The closer the judgment of the other person reaches your judgment, the higher is your probability to win an amazon voucher amounting to 50 €. Please consider that the other person does not possess any prior knowledge about the judgment task and will not see the justifications in the same order as you do. Therefore, you should describe your approach for EVERY justification in as much detail and as accurately as possible; a simple classification of the bug as toxic or not toxic will not suffice. Describe which information you used for your evaluation and how they led to the judgment. However, note that you should NOT state your judgment in the justification but only the steps towards the judgment. If your judgment is anticipated in the written justifications, you will not participate in the lottery of the Amazon voucher. Reason and justify thus properly."

[4] German versions are here, in Appendix B.

Table 3: Training items in Study 1 (multiplicative criterion) and Study 2 (linear criterion). The judgment criterion y was derived from Equation 1 (Study 1) and Equation 3 (Study 2).

Cue values				Criterion y	
x_1	x_2	x_3	x_4	Study 1	Study 2
2	1	0	3	2	14
1	4	1	4	5	22
0	3	1	2	2	13
0	2	3	0	1	12
5	5	4	0	29	43
0	4	5	4	12	26
2	4	3	0	9	26
1	4	3	5	13	27
1	0	2	4	1	12
1	0	0	2	1	6
5	3	3	5	21	40
1	1	5	5	7	22
1	2	0	5	2	15
5	5	0	1	4	36
0	4	3	1	4	19
4	2	1	3	6	27
0	5	2	3	6	22
5	5	2	4	22	43
5	1	3	4	9	33
4	0	2	4	3	24
1	4	1	5	6	23
3	0	5	5	3	27
0	2	5	0	2	16
1	5	2	4	10	27
3	4	5	5	30	39

Instructions in the verbalization condition

"In this task, it is of particular importance that you not only make a judgment, but are also able to explain what these judgments comprise. For this reason, we will randomly prompt you after some of your judgments to enter for each individual cue of the bug how many ml of toxin this cue contributed to the total toxicity of the bug. Your judgment of the total toxicity should thus result from the ml toxin that each individual cue contributes. If you can accurately state how much toxin each individual cue contributes, you will earn 20 points additionally. To do so, click on the box left to each cue with the mouse, enter the value, and confirm your response with ENTER. Enter a value for each cue."

Table 4: Validation items in Study 1 (multiplicative criterion) and Study 2 (linear criterion). The judgment criterion y was derived from Equation 1 (Study 1) and Equation 3 (Study 2).

Cue values				Criterion y	
x_1	x_2	x_3	x_4	Study 1	Study 2
3	5	1	4	10	33
3	4	4	3	21	35
5	0	3	4	4	30
3	4	2	5	14	33
5	0	5	5	4	35
3	2	0	2	2	20
2	3	4	0	9	25
4	5	4	5	36	44
5	0	5	3	4	33
4	3	0	1	3	26
2	1	2	0	3	15
2	5	2	3	12	30
4	0	0	2	2	18
4	1	1	1	4	22
3	3	3	5	15	32

Instructions for the confidence ratings

"In addition, you will be asked after each bug how much you think the response you provided deviates from the real toxicity of the bug. For instance, if you estimated 17 mg/l, but consider it possible that the toxicity of the bug ranges between 15 ang 19 mg/l, enter 2 mg/l as the response, because both 15 mg/l and 19 mg/l deviate from your estimate by 2 mg/l."

Appendix C: Cognitive modeling of judgment strategies

We followed the same cognitive modeling approach as in Hoffmann et al. (2014) to characterize participants' judgment strategies in both experiments. For each participant, we described and predicted participants' judgments with three judgment strategies: a cue abstraction strategy modeled by a linear regression model, an exemplar-based strategy modeled by an exemplar model and a guessing strategy (estimating participants' average judgment).

Cue abstraction strategies have been predominantly captured by linear regression models (Juslin, Olsson, & Olsson, 2003; Cooksey, 1996). The cue weights w_i reflect how important each cue i is for making a judgment. The final judgment $\hat{j}_p$ for an object p is determined as the sum of the

Table 5: Model Fits in the Last Three Training Blocks and in Test for Each Strategy (Guessing, Cue Abstraction, or Exemplar) in Experiment 1 (Configural Task) and Experiment 2 (Elemental Task). Standard Deviations in Parentheses.

	Experiment 1			Experiment 2	
	Justification	Verbalization	Control	Justification	Control
Model Fit Training					
Guessing	7.5 (1.5)	7.3 (1.1)	7.3 (1.0)	9.5 (1.0)	9.4 (1.0)
Cue abstraction	4.4 (0.8)	4.4 (0.9)	4.4 (0.7)	4.2 (1.5)	4.1 (1.3)
Exemplar	4.8 (3.3)	4.5 (1.3)	4.3 (1.1)	5.0 (1.7)	4.8 (1.6)
Model Fit Test					
Guessing	7.8 (2.1)	7.2 (1.4)	7.7 (1.7)	8.8 (1.7)	8.6 (1.4)
Cue abstraction	5.3 (1.5)	5.2 (1.0)	5.6 (1.4)	5.5 (1.9)	5.1 (1.6)
Exemplar	6.1 (3.1)	4.9 (1.8)	5.3 (2.1)	7.2 (2.5)	6.5 (2.1)

Note: Model fit was measured in root mean square deviation between participants' judgments and the model-predicted judgments.

cue values x_{pi} over all cues I weighted by their importance

$$\hat{j}_p = k + \sum_{i=1}^{I} w_i \cdot x_{pi} \tag{4}$$

where k is a constant intercept.

The exemplar strategy assumes that judging a new object relies upon a similarity-based retrieval of the criterion values associated with each exemplar. To model exemplar-based retrieval, we used an exemplar model with one free sensitivity parameter (Juslin et al., 2003). The similarity $S\,(p, q)$ between probe p and exemplar q is an exponential function of the objects' distance d_{pq} (Nosofsky & Zaki, 1998):

$$S(p, q) = \mathrm{e}^{-d_{pq}} \tag{5}$$

This distance is determined by summing up the absolute differences between the cue values x_{pi} of the probe and the cue values x_{qi} of the exemplar on each cue i and then weighting this sum over all cues I by the sensitivity parameter h:

$$d_{pq} = h(\sum_{i=1}^{I} |x_{pi} - x_{qi}|) \tag{6}$$

Correspondingly, the more closely the cue values of the probe and the exemplar match, the smaller the distance is between the objects. The sensitivity parameter expresses how strongly people discriminate among the stored exemplars. A sensitivity parameter close to 0 indicates no discrimination; a high parameter indicates that people specifically remember each exemplar. The estimated judgment $\hat{j}_p$ is then determined as the average sum of the similarities weighted by their corresponding criterion values y_q over all exemplars Q,

$$\hat{j}_p = \frac{\sum_{q=1}^{Q} S(p, q) \cdot y_q}{\sum_{q=1}^{Q} S(p, q)} \tag{7}$$

We estimated each model's parameters based on participants' judgments in the last three training blocks by minimizing the RMSD between participants' judgments and the model-predicted judgments and used the parameter estimates to predict participants' judgments in the four test blocks. This generalization test accounts for model complexity not only in terms of the number of free parameters but also in terms of their functional form (Busemeyer & Wang, 2000). The items for this generalization test were selected in advance to discriminate between the models (Hoffmann et al., 2014).

Descriptively, the cue abstraction strategy and the exemplar model described and predicted participants' judgments on average better than the guessing model in both experiments (see Table 5). In the configural task, the exemplar model described participants' judgments as well as the cue abstraction strategy at the end of training, but predicted participants' judgments slightly better than the cue abstraction model in the verbalization and the control condition in the test phase. In the elemental task, the cue abstraction strategy more accurately described participants' judgments at the end of training and also predicted participants' judgments better in the test phase.

Appendix D: Coding of justifications

After data collection, we asked two raters to judge the quality of participants' justifications on a range of dimensions. Each

Table 6: Means and Standard Deviations (in Parentheses) for Rated Justifications.

	Experiment 1		Experiment 2		Interrater reliability
	Mean	SD	Mean	SD	Cohen's κ
Participants	49	—	55	—	
Cues (average n mentioned)	2.6	1.0	3.0	0.9	0.91
Toxicity (% of trials)	59.0	41.0	35.3	36.4	0.81
Direction (% of trials)	43.1	34.0	33.5	36.5	0.63
Weighing (% of trials)	8.6	18.4	14.7	27.3	0.72
Combination (% of trials)	51.8	37.7	39.8	39.4	0.51
Calculations (% of trials)	0.8	4.5	7.8	25.4	0.54
Exception (% of trials)	5.3	12.3	1.8	4.7	0.29
Earlier Bugs (% of trials)	6.5	14.7	1.8	4.7	0.78
Metacognitive thoughts (% of trials)	3.9	14.6	0.7	2.6	0.52
Reason stated (% of trials)	70.2	38.1	56.2	40.8	0.84
Details (average)	3.3	0.9	3.1	1.1	*0.79
Imagery (average)	2.1	0.8	2.0	0.6	*0.73
Quality (average)	3.4	1.3	3.0	1.4	*0.81

Note: *Values represent Pearson correlations.

rater coded 10 justifications from each participant, with one justification randomly drawn from each block, without knowing from which experiment and participant the justification originated. The first four ratings involved descriptive aspects asking how many cues participants mentioned (0 to 4 cues), whether participants mentioned the overall toxicity level (binary), the direction of the relationship between the cues and the toxicity level (binary), and the importance of the cues (binary). Next, four questions were designed to better capture strategic aspects asking whether participants mentioned that combining several cues was more important than the single cues (binary), whether participants explained a way to calculate their judgment (binary), if participants mentioned that the specific bug represented an exception (binary), and if they mentioned any previously encountered bugs in their justification (binary). Furthermore, the raters made two global binary judgments involving if the justification included any metacognitive thoughts and if the description actually comprised any reasons for the judgment. The binary rating if the justification comprised any reasons was used to discriminate pure descriptions of the bug from justifications stating reasons why a bug is more or less toxic. Although participants often described the bug instead of providing reasons, they rarely entered no justification at all. In addition, raters made three global judgments on a Likert scale asking how detailed the description was (7-point Likert scale from 1 = "no details" to 7 = "many details") or how figurative the description was (7-point Likert scale from 1 = "prosaic" to 7 = "figurative"), and finally raters judged the overall quality of the justification by considering how helpful the justification was for deriving a judgment (7-point Likert scale from 1 = "useless" to 7 = "very helpful"). Example justifications highlighted typical statements representing each category.

Interrater reliability was satisfying for most descriptive aspects of the ratings, but lower for questions capturing strategic aspects (Table 6 summarizes interrater reliability and descriptive statistics). In particular, ratings did not agree on the classifications of exceptions, potentially because participants mentioned only vaguely in their justifications that the object under consideration has to be judged differently than all the other ones. In case of such conflicts, a third rater judged the justifications again. The last three global ratings (Details, Imagery, and Quality) were averaged across the two raters. Finally, we normalized all ratings to a range between 0 and 1 and summarized the four descriptive questions (Cues, Toxicity, Direction, Weighting) and the global quality rating within a quality index ranging from 0 (justifications did not include any information) to 1 (justifications included information about the cues, the toxicity, the weighting, the direction, and a global quality rating).

Among those justifications for which participants stated reasons participants primarily mentioned that combining several cues was important and slightly considered a combination of cues more often in the configural task from Experiment 1 (*M* = 62.0%, *SD* = 35.5%) than in the elemental task from Experiment 2 (*M* = 57.9%, *SD* = 40.7%). Stating a rule for calculating the judgment seldom happened, but was slightly more pronounced in the elemental task (*M* = 9.8%, *SD* = 28.3%) than in the configural one (*M* = 0.9%, *SD* = 5.0%). Finally, participants rarely mentioned earlier bugs or exceptions in their justifications, nor did the percentage of references vary between the configural task (Bugs: *M* = 7.3%, *SD* = 15.4%, Exceptions: *M* = 6.6%, *SD* = 14.9%) and the elemental task (Bugs: *M* = 5.7%, *SD* = 17.4%, Exceptions: *M* = 4.6%, *SD* = 15.8%).

Similarly, which reasons participants stated did not strongly differ between participants classified to the cue abstraction model (or the exemplar model) across both experiments. Participants classified to the cue abstraction model slightly mentioned a combination of cues more often (*M* = 59.7%, *SD* = 37.5%) as well as a way to calculate the judgment (*M* = 6.8%, *SD* = 24.7%), but referred less often to previous bugs (*M* = 6.4%, *SD* = 16.5%) or mentioned exceptions (*M* = 4.5%, *SD* = 14.3%). In contrast, participants classified to the exemplar model more often considered previous bugs (*M* = 7.7%, *SD* = 18.0%) or mentioned exceptions (*M* = 8.0%, *SD* = 18.4%), but less often stated they calculated the judgment (*M* = 3.7%, *SD* = 12.8%) or that a combination of cues was important (*M* = 55.6%, *SD* = 41.7%).