Graduate School of Decision Sciences

Universität Konstanz

GSDS
Working Paper
No. 2017-04

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

Janina A. Hoffmann
Wolfgang Gaissmaier
Bettina von Helversen

February 2017

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
February 2017

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

Janina A. Hoffmann
University of Basel
University of Konstanz

Wolfgang Gaissmaier
MPI for Human Development
University of Konstanz

Bettina von Helversen
University of Basel
University of Zurich

The quality of a decision in daily life is often evaluated based on whether one can adequately explain how one came to this decision. Holding decision makers accountable for the decision process often improves judgment quality because decision makers weigh and integrate information more thoroughly. Our research aimed to identify the conditions under which process accountability may lead to favorable or disadvantageous judgments. Specifically, we hypothesized that process accountability may hinder accurate judgments in tasks that require remembering previously encountered cases (Experiment 1) whereas it may improve judgment accuracy in tasks that require weighing and integrating information (Experiment 2). To test these hypothesis, participants learned to solve either a multiplicative (Experiment 1) or a linear multiple-cue judgment task (Experiment 2) with or without accountability instructions. We manipulated process accountability by randomly selecting trials in which participants had to justify their judgment. In both experiments, process accountability neither changed how accurately people made a judgment, nor the cognitive processes underlying their judgments. In trials asking participants to justify their judgment process participants were less confident about their judgments. Overall, these results imply that process accountability may impact judgment quality only to a small. We discuss the limits of our results with respect to the effectiveness of the manipulation.

Keywords: Judgment; Accountability; Cognitive processes

Introduction

The idea that the quality of decisions can be improved by holding decision makers accountable for their judgments is wide spread. Indeed, 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 accountability and process accountability (Langhe, van Osselaer, & Wierenga, 2011; Siegel-Jacobs & Yates, 1996). If decision makers are held accountable for the outcome, their performance is solely evaluated based upon the outcome of their decisions. For instance, investors may sell stocks of a low-performing company regardless of whether managerial decisions are to blame for the financial underperformance. In contrast, if decision makers are held accountable for the process, their performance is evaluated based on how they formed a judgment. Legally, for instance, a bad business decision is one that violates due care, but not necessarily one that leads to a financially bad outcome.

Even though one might expect that being accountable for the outcome of one's judgments increases judgment accuracy more than just being accountable for the process, research suggests the opposite. Whereas outcome accountability has been shown to have mostly negative side-effects, process accountability has been found to benefit judgment per-

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

Correspondence concerning this article should be addressed to Janina A. Hoffmann, Department of Psychology, University of Konstanz, UniversitÃďtsstrasse 10, 78 468 Konstanz, Germany. E-mail: janina.hoffmann@uni-konstanz.de

formance in a wide range of tasks (DeCaro, Thomas, Albert, & Beilock, 2011; Lerner & Tetlock, 1999). One reason for these benefits is potentially that people pay more attention to the judgment problem if they expect to justify their judgment process. As a consequence, they may use information more consistently or store and retrieve available information more accurately (Langhe et al., 2011). This thorough processing, in turn, may increase the accuracy and consistency of judgments and attenuate cognitive biases (Langhe et al., 2011; Lerner & Tetlock, 1999).

Yet, the advantages of process accountability may be limited to situations in which people weigh and integrate all available pieces of information. Research indicates that persons imagining that they are held accountable for the judgment process prefer to consider more information and to trade-off this information (Kahn & Baron, 1995). Indeed, although participants who are held accountable for the judgment process often consider more relevant cues, they are also more likely to consider more irrelevant information (Siegel-Jacobs & Yates, 1996). Similarly, Langhe et al. (2011) found that performance increased only in multiple-cue judgment tasks in which people deliberately weigh and integrate information in a rule-based fashion, but not in tasks that are best solved by retrieving past instances from memory (exemplar-based strategies). Taken together, process accountability may benefit performance the most in tasks demanding weighing and integrating information. However, the reasons why tasks eliciting exemplar-based strategies do not profit from process accountability still remain unclear. The goal of the current paper is to better understand when and why process accountability may help or hurt judgment performance by taking a closer look at how process accountability can influence the cognitive strategies people rely on when making judgments. In the following we will summarize research on multiple-cue judgment tasks and the cognitive strategies people use to solve them. Then we will discuss how process accountability may influence the cognitive processes and ultimately judgment performance. Finally we report two experiments in which we investigated these questions empirically.

Judgment strategies in multiple-cue judgment tasks

Multiple-cue judgments refer to decision problems in which the judge has to evaluate a person or object on a continuous criterion based on a number of attributes or cues such as a teacher grading a student essay. When judging the essay the teacher determines the grade (the criterion) based on indicators of the essays' quality (i.e. the cues) such as the clarity of the writing style, the coherence of the reasoning, or the adherence to formal criteria such as citation rules.

Over the past decade, evidence has accumulated that people base their judgments on different judgment strategies. In particular two types of strategies have been proposed: cue abstraction and exemplar memory (Hoffmann, von Helversen, & Rieskamp, 2014; Juslin, Karlsson, & Olsson, 2008; von Helversen & Rieskamp, 2008). Cue abstraction strategies assume that people try to understand how each cue relates to the criterion and then make a judgment by linear additively integrating them. That is people weigh each cue by its importance and then integrate them to a final judgment. For instance, new teachers may learn to give more importance to the coherence of the reasoning than formal criteria when grading students' essays.

In contrast, exemplar-based strategies assume that people retrieve information about previously stored exemplars when judging a new instance. The higher the similarity is of a stored exemplar to the to-be judged object, the more this past exemplar influences the final judgment. For instance, tutors could try to judge students' essays based on a number of example cases they received from the professor.

Past research suggests that people have both strategies at their disposal. Which strategy is selected for the current task depends on the characteristics of the task and the cognitive abilities of the decision maker (Hoffmann, von Helversen, & Rieskamp, 2013, 2014; Juslin et al., 2008; Mata, von Helversen, Karlsson, & Cüpper, 2012; von Helversen, Mata, & Olsson, 2010). Specifically, in linear or elemental judgment tasks in which the criterion is a linear function of the cues people's judgments are well described by the cue abstraction strategy (Hoffmann, von Helversen, & Rieskamp, 2016; Juslin et al., 2008). In contrast, in non-linear or configural judgments tasks in which the criterion is non-linear function of the cues, people tend to rely on exemplar memory —possibly because ceu abstraction strategies are less successful in these tasks (Hoffmann et al., 2013; Juslin et al., 2008). Furthermore, how accurately people solve these tasks has been linked to different memory abilities of the decision maker. Whereas performance in an elemental task depends more on working memory capacity, performance in a configural tasks depends more on episodic memory (Hoffmann et al., 2014).

Effects of process accountability on judgment strategies

How should process accountability interact with judgment strategies? Langhe et al. (2011) argued that process accountability particularly boosts cue abstraction, but leaves exemplar memory unaffected. Specifically, they reasoned that process accountability may increase epistemic motivation, the motivation to acquire a thorough understanding of the decision process (De Dreu, Beersma, Stroebe, & Euwema, 2006; Langhe et al., 2011), and thus encourage a more systematic processing of available information. This increase in systematic processing 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, but not in a configural judgment task. Furthermore, modelling the judgments strategies in the elemental judgment task revealed that how consistently participants were described by cue abstraction strategies mediated the beneficial effect of process accountability on judgment performance for participants low in rationality. In sum, these results indicate that process-accountability motivated participants to execute a cue abstraction strategy more accurately and more consistently.

It still remains unclear, however, how process accountability may affect exemplar memory. Langhe et al. (2011) reasoned that exemplar-based processes are fairly automatic and thus processing the available information more systematically may not be of any help if people base judgments on exemplar memory. Consistent with this idea, Langhe et al. (2011) found no differences in judgment accuracy between process and outcome accountability in a non-linear, quadratic task, but did not investigate which cognitive strategies people used. Modeling the cognitive strategies in this quadratic task may have been particularly informative, because current research still debates if people solve the task by storing exemplars or if they rather drop back to an unsuccessful cue abstraction strategy (Hoffmann et al., 2016; Olsson, Enkvist, & Juslin, 2006; Pachur & Olsson, 2012). Specifically, both types of strategies do not allow a high performance early in training and participants may detect the appropriate strategy only after a long training phase (Olsson et al., 2006; Pachur & Olsson, 2012). Accordingly, although process accountability may have affected the degree to which participants engaged in exemplar memory or cue abstraction, changes in the cognitive process may not have translated to judgment performance in the quadratic task. 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 as in the elemental task.

If the latter view holds, process accountability may disrupt judgment performance in tasks that are usually solved by exemplar-based strategies. Exemplar-based strategies are often assumed to be intuitive strategies that are difficult to verbalize (XX;XX). The necessity to explain ones' judgment process may force participants to verbalize their strategy and, in turn, introduce a processing shift that interferes with non-verbal processes such as exemplar memory (Schooler, 2002). In this vein, concurrent verbalization has been shown to impair accuracy on Raven's Matrices, particularly for items dependent on visuospatial processes, (Deshon, Chan, & Weissbein, 1995). Similarly, evidence from categorization research suggests that process accountability selectively interferes with procedural categorization strategies that should be harmed by awareness of the judgment process (DeCaro et al., 2011; McCoy, Hutchinson, Hawthorne, Cosley, & Ell, 2014). For instance, it has been found that videotaping participants during category learning tasks and announcing that their performance will be watched by others hurts performance in information-integration tasks, but not in rule-based category tasks (DeCaro et al., 2011). One reason for this performance decrease is that participants were more likely to consider two- and three-dimensional rules and less often a procedural information integration strategy. Thus, if process accountability causes people to switch away from an exemplar-based strategy to a rule-based strategy, it may harm performance in tasks that are better solved by an exemplar-based strategy.

To foreshadow our results, we do not find much support fot the hypothesis that process accountable participants made less accurate judgments in an exemplar-based task. Although process accountable participants indicated a lower confidence in their judgment, not a higher percentage of participants was better described with a cue abstraction strategy, nor did people make more consistent judgments.

Experiment 1: Accountability in a multiplicative judgment task

To test the prediction that process accountability impairs judgment accuracy in exemplar-based judgment tasks, we manipulated the need to justify one's judgment process while participants learned to solve a multiple-cue judgment task. Similarly to Langhe et al. (2011) we chose a configural task, but selected a multiplicative task that (a) more reliably induces exemplar-based processes and (b) allows identifying the strategies people used (Hoffmann et al., 2014, 2016). We aimed at pinning down the cognitive strategies specifically involved when participants have to justify their judgment process. Accordingly, in contrast to Langhe et al. (2011) we compared the process accountability condition to two control conditions: One without an accountability instruction to investigate the degree to which process accountability impairs judgment accuracy and one with a verbalization instruction to identify the degree to which justifying one's judgment process towards others impedes judgment accuracy more than just indicating how much each cue contributed to the final judgment. We introduced this verbalization condition because it has been suggested that a mere verbalization may also interfere with non-verbal processes (Deshon et al., 1995; Schooler, 2002; Schooler & Engstler-Schooler, 1990). In the accountability and the verbalization condition participants were repeatedly prompted to justify (or verbalize) their judgment process after a random sample of trials, twenty times over the course of learning. Informing participants that they have to justify their judgment process after a random sample of trials should increase the motivation to explicitly reason about the judgment process in each single trial more than informing participants that they have to uncover their judgment process only at the end (cf. Langhe et al., 2011). Further, providing a justification directly after the judgment reduces retrospection which may distort the validity of the justification, particularly if judgment strategies change during the course of learning (Lagnado, Newell, Kahan, & Shanks, 2006). Finally, we asked participants to provide confidence ratings after every trial allowing us to assess how justifying one's judgment process affects confidence in one's judgment.

Method

Participants. Hundred fifty three participants were recruited from the participant pool of the Max-Planck-Institute for Human Development, Berlin. Due to an error in the experiment, we discarded 9 incomplete data sets, leaving a sample of 144 participants (80 female, $M_{\text{Age}} = 25.4$, $SD_{\text{Age}} = 3.3$). Participants received an hourly wage of XX €for their participation as well as a performance-dependent bonus ($M = 2.9$ €, $SD = 0.84$ €).

Design and Material. We adapted a judgment task from Hoffmann et al. (2016) asking participants to estimate the toxicity of a bug (the criterion) on a scale from 0 to 50 mg/l. This judgment criterion was predicted by four quantitative cues with cue values ranging from 0 to 5. The function generating the judgment criterion y included two multiplicative combinations of the cues x_1,..., x_4:

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

We used the same items in the judgment task as in previous studies (Hoffmann et al., 2014, 2016). Table 1 depicts the training set and Table 2 the validation set for Experiment 1.

Table 1
Training Items in Study 1 (multiplicative criterion) and Study 2 (linear criterion).

Cue values				Criterion	
C_1	C_2	C_3	C_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

The stimuli consisted of pictures of 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 toxicity of the bug. For each participant, the cues x_1,..., x_4 were randomly assigned to the visual features of the bug (e.g., antennae). Higher cue values, however, were always associated with more salient visual features. For instance, a cue value of zero corresponded to a bug without spots on the back and a cue value of five to a bug with five spots on its back. Likewise, 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.

Procedure. In the beginning of the experiment, all participants were instructed that they will learn to predict the toxicity of different bugs during the training phase. In the justification condition, participants were additionally informed that they will need to justify their judgments later so that another person could make the same judgments based upon their descriptions. In the verbalization condition, participants were informed that they will be asked later to subdivide their judgments into its components.

To facilitate imagining which information they (or another person in the justification condition) would need to accurately judge the bugs' toxicity, all participants first solved a short practice task. In this practice task, participants were asked to imagine that they had to judge the toxicity only based upon the bugs' description. They next saw a bug with different cues and were asked to type in which information and knowledge they would need to make an accurate judgment.

Afterwards, participants moved on to the judgment task that consisted of a training phase and a test phase. During the training phase, participants learned to estimate the criterion values for 25 training items from the training set. In each trial, participants first saw a picture of a bug and were asked to estimate its toxicity (the criterion value). Afterwards, participants judged how confident they were about their judgments by estimating how much their own answer deviated from the correct judgment. At the end of each trial, they received feedback about the correct value, their own estimate, and the points they had earned. The training phase ended after 10 training blocks, each consisting of the 25 training items presented in a random sequence.

In 20 of these 250 judgment trials, the experimental trials,

Table 2
Validation items in Study 1 (multiplicative criterion) and Study 2 (linear criterion).

Cue values				Criterion	
C_1	C_2	C_3	C_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

participants who had to justify their judgment were asked 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 were asked to indicate how much each cue contributed to the total amount of toxicity. Verbalization questions and justifications occurred randomly two times in each training block directly after participants made their judgment. Thus participants could not know beforehand in which trials they would need to justify (or verbalize) their judgment. The sequence of each trial is depicted in Figure 1.

In the subsequent test phase, participants judged 15 new validation items four times and indicated their confidence but did not receive any performance feedback. Further, participants did neither answer any verbalization questions nor were they asked to justify their judgments.

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

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

At the end of the judgment tasks, 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 a root mean square deviation [RMSD] of less than 5.5 in both judgment tasks). Finally, we also incentivized participants in the verbalization and the justification condition for carefully responding to the verbalization questions (or to carefully explain their judgment, respectively). Participants in the verbalization condition could gain 20 additional points for each verbalization question if the importance they assigned to each cue summed up to their own judgment. Participants in the justification condition could win one Amazon voucher worth 50 €. Chances of winning the voucher were higher for a participant the more closely another person could approximate his or her judgment based on the participants' justification and the bug the participant saw. [1]

Results

Does justification decrease judgment performance?. Across all conditions, participants on averaged learned to solve the judgment task well. As a measure of judgment error in each training block, we calculated the RMSD between participants' judgments and the correct criterion. Descriptively, judgment error dropped in all conditions from the first training block to the last training block and participants made slightly worse judgments in justification (see Table 3 for descriptive statistics and Figure 2 for mean performance in each block). Furthermore, the majority of participants reached the learning criterion in all conditions. In the justification condition, 39 out of 49 participants earned a bonus (79.6%, yet 4 participants did not outperform a guessing model). Similarly, 80.9% of the participants in the verbalization condition (38 out of 47, with 3 participants worse than guessing) earned a bonus as well as 87.5% in the control condition (42 out of 48, with 2 participants worse than guessing). How many participants earned a bonus in the judgment task, did not vary across experimental conditions, χ^2 (2) = 1.21, p = 0.545, Bayes Factor (BF) = 0.058 [2], indicating that participants in all conditions learned the task comparably well.

To investigate more closely if judgment performance changed over the course of learning as a function of justification, we conducted a repeated measures ANOVA on judg-

[1] To measure how closely another person approximated the judgment of the participant, we randomly selected five justifications for each participant in the justification condition, resulting in 320 justifications in total. If participants stated their judgment in these justifications, the judgment was replaced by "XX". In a later rating study, a rater was presented in each trial with one justification as well as with the corresponding bug. Based upon this information, the rater was asked to approximate the judgment. All justifications were randomly interspersed and the rater was aware that those justifications were generated by different participants.

[2] BFs were calculated using the BayesFactor Package (Morey, Rouder, Jamil, & Morey, 2015) and indicate the extent to which the data supports the alternative hypothesis over the null hypothesis. BFs above 10 provide strong evidence for the alternative hypothesis, BFs below 1/10 provide strong evidence for the null hypothesis(Jeffreys, 1961).

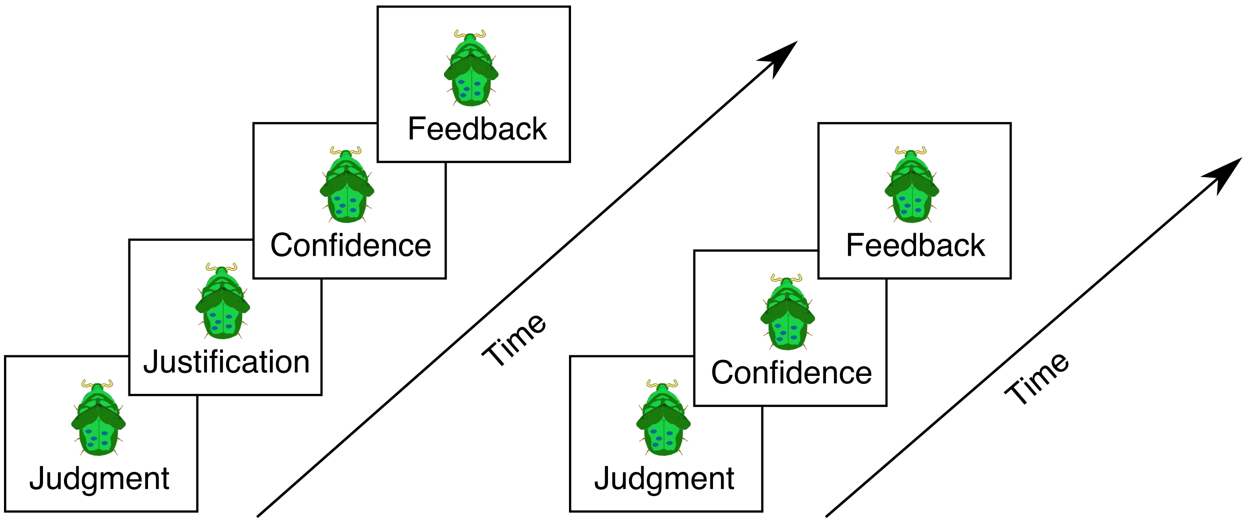

Figure 1. Trial sequence for experimental (left sequence) and control trials (right sequence). In the experimental trials, participants in the justification condition had to justify their judgment after they made a judgment, whereas participants in the verbalization condition indicated how much each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

Table 3
Performance in Experiment 1 (Multiplicative Environment) and Experiment 2 (Linear Environment). 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)
Test session					
Mean Error	6.2 (3.0)	5.7 (1.8)	5.5 (1.9)	6.2 (1.9)	5.8 (1.7)
z-Confidence					
Pre-Trial	0.02 (0.71)	-0.01 (0.71)	-0.03 (0.55)	0.06 (0.68)	-0.09 (0.5)
Trial	0.21 (0.9)	-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)
z-Accuracy					
Pre-Trial	0.06 (0.7)	-0.04 (0.23)	-0.02 (0.26)	0.05 (0.4)	0 (0.4)
Trial	0.07 (0.75)	-0.02 (0.3)	-0.01 (0.26)	0.09 (0.37)	-0.07 (0.32)
Post-Trial	0.03 (0.58)	-0.05 (0.33)	-0.04 (0.33)	0.01 (0.42)	-0.07 (0.33)

Note. Error in the judgment tasks was measured as the Root Mean Squared Deviation.

ment error in each training block (Greenhouse-Geisser corrected) with the independent factors training block and condition (justification, verbalization, and control). This analysis indicated that participants made less errors in later training blocks, $F_{\text{Block}}(5.11,720.01) = 167.54$, $p < 0.001$, $\eta^2 = 0.22$,$\text{BF}_{\text{Block},0} > 0.058$. However, participants who justified or verbalized their judgment did not make more errors than participants in the control group, $F_{\text{Cond}}(2,141) = 0.92$, $p = 0.400$, $\eta^2 < 0.01$, nor did the need to justify or verbalize ones' judgment change judgment performance over time, $F_{\text{Block x Cond}}(10.21,720.01) = 1.72$, $p = 0.070$, $\eta^2 = 0.01$. Similarly, BFs preferred a model indicating that judgment error decreased over time ($\text{BF}_{\text{Block},0} > 10000$), but this model was not outperformed by a model assuming that justification affected judgment error ($\text{BF}_{\text{Cond+Block,Block}} = 0.269$) or by a model assuming an interaction between condition and block($BF_{\text{Block x Cond,Block}}$ = 0.036). Also a more specific test restricting the order of conditions (Justification > Insight > Control) provided only evidence hardly worth mentioning for the hypothesis that participants were less accurate in justification than in the verbalization condition and far less accurate than participants in the control condition (BF = 2.591).

As in training, we measured judgment error in the test blocks as the RMSD between the correct criterion and participants' judgments. Average judgment error across the four test blocks was slightly higher in justification than in the verbalization or the control condition. A repeated measures ANOVA on judgment error in the four test blocks suggested that participants were less accurate in later test blocks, $F_{\text{Block}}(2.96, 416.85) = 2.87$, $p = 0.037$, $\eta^2 < 0.01$, but a corresponding BF analysis attributed this main effect to the random effect for participants ($BF_{\text{Block,0}} = 0.318$). As in training, participants who had to justify or verbalize their judgments did not make less accurate judgments than participants in the control group, $F_{\text{Cond}}(2,141) = 1.15$, $p = 0.319$, $\eta^2 < 0.01$, nor did the influence of justification on judgment error change over blocks, $F_{\text{Block x Cond}}(5.91,416.85) = 1.28$, $p = 0.267$, $\eta^2 = 0.01$. Similarly, BF's did not favor a model assuming that justification affected judgment errors over the nul model, $BF_{\text{Cond,0}} = 0.45$, nor a model assuming an interaction between condition and block, $BF_{\text{Cond x Block,0}} = 0.006$. Setting ordered constraints neither provided support for the idea that participants in justification judged test items less accurately than participants in the verbalization and the control condition (BF = 2.647). In sum, participants learned to make accurate judgments with more training blocks and kept a high judgment performance in test, but neither justification nor deconstructing their judgment decreased judgment accuracy in training or test.

Judgment strategy, accuracy and consistency. To better understand on which judgment strategy participants based their judgment, we fitted three cognitive judgment models to participant's judgments in training and predicted their judgments in test (cf. Hoffmann et al., 2014, 2016): a cue abstraction model, an exemplar model, and a guessing model that estimated participant's mean judgment. Overall, the majority of participants was best described by an exemplar model and only one participant in justification was best described by guessing (see 4 for strategy classification, average test performance and strategy consistency by strategy). The number of participants best described by a cue abstraction model did not vary between conditions, simulated $\chi^2 = 3.8$, $p = 0.407$, BF = 0.002. In justification, the cue abstraction model predicted judgments of 49 % of participants best, whereas it predicted judgments 36.2 % of participants in the verbalization and of (43.8 %) of participants in the control condition best. To assess to what degree the judgment strategy people chose affected judgment performance in test, we included judgment strategy in the repeated measures ANOVA on judgment error in each test block with block, judgment strategy, and condition as independent variables, excluding participants best described by guessing. Overall, participants assigned to the exemplar model were more accurate in test, $F_{\text{Strategy}}(1, 137) = 21.31$, $p = 0.001$, $\eta^2 < 0.12$, $BF_{\text{Strategy,0}} = 1967$, but people assigned to a cue abstraction model in justification or verbalization did not make more errors in test than the participants assigned to the cue abstraction model in the control group, $F_{\text{Strategy x Cond}}(2, 137) = 0.78$, $p = 0.460$, $\eta^2 < 0.01$, $BF_{\text{Strategy x Cond,Strategy}} = 0.136$. All other interactions were not significant.

In addition, we investigated whether justification and strategy use influenced how consistently participants judged the same items across the test blocks. To account for this possibility, we calculated the consistency for each participant as the average correlation between the judgments in the test blocks. All analysis were performed using Fisher's z-transformed correlations. Participants who were best described by a guessing strategy were excluded from this analysis. Descriptively, participants best described by a cue abstraction model made less consistent judgments across the test blocks compared to participants best described by an exemplar model, but consistency did not vary systematically depending on justification. An ANOVA on consistency similarly indicated that participants classified to the cue abstraction model made less consistent judgment in test than participants classified to the exemplar model, $F_{\text{Strategy}}(1, 137) = 8.92$, $p = 0.003$, $\eta^2 < 0.06$, $BF_{\text{Strategy,0}} = 8.227$, but participants in justification did not make more consistent judgments than participant in the verbalization or control condition, $F_{\text{Cond}}(2, 137) = 0.12$, $p = 0.886$, $\eta^2 < 0.01$, $BF_{\text{Cond,0}} = 0.069$. Finally, justifying one's judgments did not more strongly affect strategy execution for participants classified to the cue abstraction model than for participants classified to the exemplar model, $F_{\text{Strategy x Cond}}(2, 137) = 1.63$, $p = 0.200$, $\eta^2 < 0.02$, $BF_{\text{Strategy x Cond,0}} = 0.231$. In sum, justification did neither lead to a shift to more rule-based judgment strategies, nor did it change how consistently participants applied a learned strategy.

Analyzing confidence in one's judgment. As a manipulation check, we also asked participants after every trial to indicate how far their judgment deviated from the correct judgment (confidence). If the need to justify one's judgment decreased participants' confidence in their own judgment, they should have indicated that they further deviated from the correct criterion in trials in which they had to justify their judgment compared to preceding or subsequent trials without this prompt. To only consider 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 question), trials

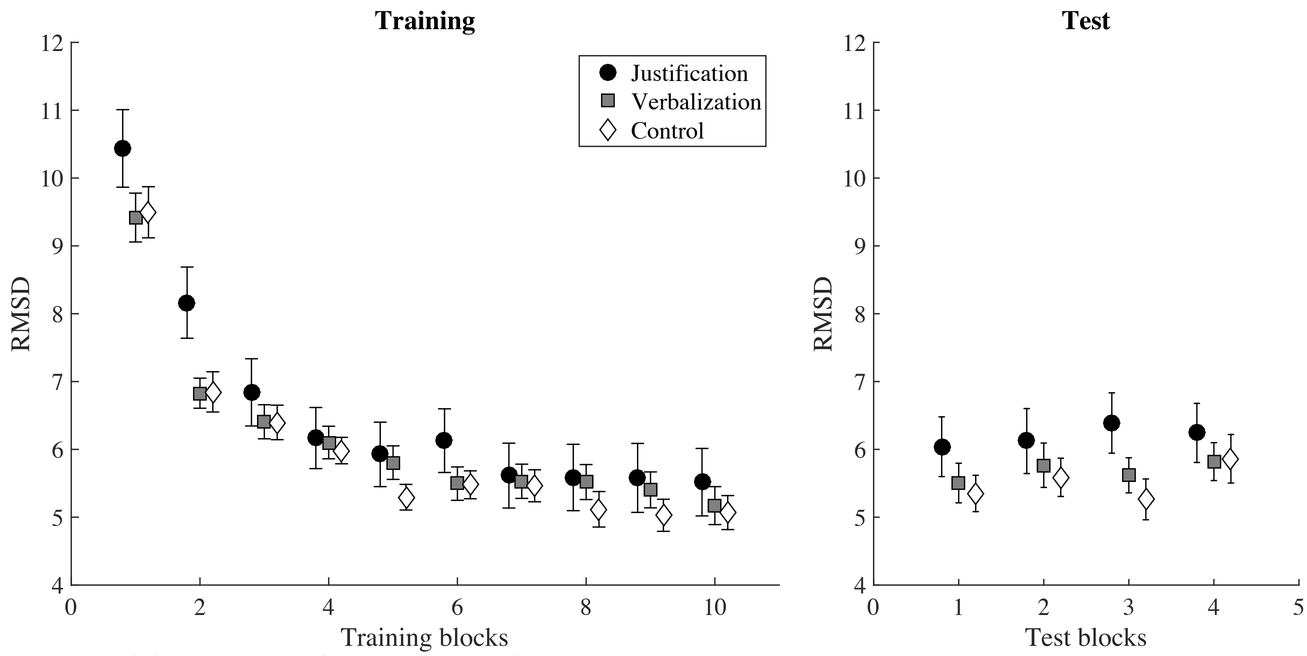

Figure 2. Training and test performance in Experiment 1.

that contain a justification, and trials after the justification. 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 3 depicts z-standardized confidence ratings in each condition, separately for each trial type. Descriptively, participants in the justification condition are slightly less confident in trials in which they had to justify their judgment, compared to the control or the verbalization condition. A repeated measures ANOVA (Greenhouse-Geisser corrected) suggested that participants in the justification condition were not generally less confident about their judgments, $F_{\text{Cond}}(2, 141) = 0.41$, $p = 0.668$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond},0} = 0.37$. However, participants were less confident in trials with a justification than in trials preceding or following a justification, $F_{\text{Trial type}}(1.58, 222.26) = 7.42$, $p = 0.002$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial type},0} = 22.6$. How strongly confidence changed as a function of trial type depended on the condition, $F_{\text{Cond x Trial type}}(3.15, 222.26) = 3.08$, $p = 0.026$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond x Trial type},0} = 0.456$. To further check if the trial type differently affected confidence in the three conditions, we set equality constraints on the factor trial type, separately for each condition. Put differently, we assumed for each condition separately that participants were as confident before justification as in the justification trial or after justification and that their confidence level equalled the confidence level of participants in the other conditions before justification. We finally compared this restricted model with the unrestricted model, with both models including the interaction. This comparison preferred the restricted model for the control (BF = 5.1) and the verbalization condition (BF = 8.8), but favored the unrestricted model for the justification condition (BF = 0.0002). In sum, this result suggests that participants who had to justify their judgment were less confident in trials including this justification, but not participants who to verbalize their judgment or did not engage in justification.

It is possible that participants who were less confident on the justification trials also make more errors in trials following the justification. To also consider to what degree accuracy changes depending on whether participants had to justify or verbalize their judgment in the trial before, we repeated the above analysis with z-standardized accuracy. On average, participants were as accurate in justification as in verbalization or the control group, $F_{\text{Cond}}(2, 141) = 0.61$, $p = 0.542$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond},0} = 0.294$. Moreover, participants were not less accurate in trials following the justification than in justification trials or trials preceding a justification, $F_{\text{Trial}}(1.89, 267.03) = 0.85$, $p = 0.423$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial},0} = 0.1$. Finally, trial type did not interact with the condition, $F_{\text{Trial x Cond}}(3.79, 267.03) = 0.04$, $p = 0.995$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial x Cond},0} = 0.004$, indicating that justification also did not influence how accurately participants made their judgment in the trial directly following the justification prompt. Setting equality constraints led to similar results. In sum, this analysis suggests that justification neither had an instantaneous effect that worsens performance in the trial directly following the justification.

Discussion

To summarize our main findings, neither process accountability nor verbalization decreased judgment accuracy in the multiplicative task compared to a control group receiving only outcome feedback. This finding resonates well with the finding that justifying the judgment process compared to justifying the outcome does not affect performance accuracy in configural, quadratic tasks (Langhe et al., 2011). We added to this research by modelling the cognitive strategies participants may have used to solve the judgment task. This strategy classification also provided no support for the idea that process accountability induced a shift to a cue abstraction strategy. In combination, these results hint at the interpretation that justifying one's judgment process does not interfere with exemplar memory and provide support for the notion that exemplar memory operates on more automatic processes (Langhe et al., 2011).

Despite the lack of influence on absolute performance we found that participants were less confident in their judgment after justification. Thus, process accountability made participants reconsider their judgment strategy, although the corresponding results on judgment accuracy suggest that this reconsideration did not extend to the subsequent trial. These results make it likely that process accountability affects judgment performance to a smaller extent than suggested in prior research. In line with this idea, Siegel-Jacobs and Yates (1996) found that holding participants accountable for the process failed to affect overall judgment accuracy compared to a condition without accountability instruction and only improved calibration (Exp. 1) or discrimination (Exp. 2). If so, the benefitial effects of process accountability in an elemental task may be also overstated. Will manipulating process accountability improve judgment accuracy in an elemental judgment task?

Experiment 2: Accountability in an elemental judgment task

In elemental judgment tasks, the benefits of process accountability over outcome accountability are well documented (XX;XX). In three experiments, Langhe et al. (2011) provided convincing evidence that justifying the judgment process improves accuracy more than justifying the outcome. Similarly, it has been found that stating reasons for one's judgment can promote a higher judgment accuracy even in the absence of social pressure (Ashton, 1990, 1992). One reason for why process accountability may be benefitial for judgment performance is that people make more consistent judgment possibly because they applay a cue abstraction strategy more consistently (Ashton, 1990, 1992; Langhe et al., 2011). Specifically, we hypothesized that holding participants accountable for the judgment process may lead to more accurate judgments in an elementad judgment task in which most participants should rely on cue abstraction. To test this prediction, we compared the effect of process accountability against a control condition without accountability in an elemental judgment task.

Method

Participants. Hundred-ten participants (58 females, M_{Age} = 25.6, SD_{Age} = 6) from the University of Basel were recruited for this experiment with 55 participants in each condition. Participants 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).

Material, Design, and Procedure. We used the same stimulus material as in Experiment 1 and only varied the function relating the cues to the criterion. Specifically, in Experiment 2, the judgment criterion y was a linear, additive combination of all cues, x_1, ..., x_4:

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

Table 1 and 2 displays the task structure in the linear environment. We did not change any procedural details besides adapting the monetary incentive to Swiss Francs. Specifically, the points earned were converted to a monetary bonus at a rate of 1500 points = 1 CHF and particpants earned an additional bonus of 3 CHF if they reached 80 % of the points in the last training block. Finally, participants in the justification condition could win one Amazon voucher worth 50 €.

Results

Does justification increase judgment performance?. As in Experiment 1, participants on averaged learned to solve the judgment task well in both conditions. Judgment error dropped in all conditions from the first training block to the last training block, but participants made slightly worse judgments in justification (see Table 3 for descriptive statistics and Figure 3 for mean performance in each block). Justifying one's judgment did not affect how many participants gained a bonus in the training phase. In the justification condition, 38 out of 55 participants earned a bonus (69.1% with 1 participant worse than guessing). Similarly, 44 out of 55 participants in the control condition reached the learning criterion (80% with 2 participants worse than guessing), χ^2 (1) = 1.2, p = 0.274, BF = 0.475.

To investigate if participants in the justification condition learned to solve the judgment task faster over time and made more accurate judgments, we measured judgment error in each training block. As in Experiment 1, a repeated measures analysis (Greenhouse-Geisser corrected) indicated that judgment error on average judgment error decreased over the course of learning, $F_{\text{Block}}(5.44,587.49)$ = 105.83, p < 0.001, η^2 = 0.25,$\text{BF}_{\text{Block},0}$ > 10000. Yet, participatns did not make more accurate judgments in justification than in the control condition, $F_{\text{Cond}}(1,108)$ = 1.83, p < 0.179, η^2 = 0.01,$\text{BF}_{\text{Block + Cond,Block}}$ > 0.56. Finally, participants who had to justify their judgments also did not show a steeper learning path than participants in the control condition, $F_{\text{Block x Cond}}(5.44,587.49) = 0.92$, $p < 0.477$, $\eta^2 = 0$,$\text{BF}_{\text{Block x Cond,Block}} > 0.003$. Setting ordered constraints (Justification < Control) likewise provided more evidence against the hypothesis that participants in justification achieve a better performance than participants in the control condition (BF = 0.221 compared to the full model).

In the test blocks, participants made descriptively more judgment errors in justification than in the control condition. Analyzing judgment error in each test block using a repeated measures ANOVA indicated that participants made more errors in later test blocks, $F_{\text{Block}}(2.62,283.35) = 2.78$, $p < 0.049$, $\eta^2 = 0.01$, but a corresponding Bayesian ANOVA did not provide evidence for this main effect,$\text{BF}_{\text{Block,0}} = 0.374$. Participants who had to justify their judgments were not more accurate than participants in the control group, $F_{\text{Cond}}(1,108) = 1.94$, $p < 0.166$, $\eta^2 = 0.01$, $\text{BF}_{\text{Cond,0}} = 0.639$. Furthermore, how accurately participants justifying their judgments were did not change over the test blocks, $F_{\text{Block x Cond}}(2.62,283.35) = 1.11$, $p < 0.342$, $\eta^2 = 0.01$, $\text{BF}_{\text{Block x Cond,0}} = 0.022$. As in training, testing ordered constraints (Justification < Control) provided more evidence against the hypothesis that participants in justification achieve a better performance than participants in the control condition (BF = 0.214 compared to the full model).

Judgment strategy and weighting. As in Experiment 1, we modeled participants' judgment strategies to gain some insight into the underlying judgment processes. In contrast to Experiment 1, however, the majority of participants was best described by a cue abstraction model (see Table 4). Only two participants in the justification condition and one participant in the control condition was best described by guessing. Yet, the number of participants best described by cue abstraction did not vary as a function of justification, simulated $\chi^2 = 1.61$, $p = 0.458$, BF = 0.056. In the control condition, 87.3% of the participants were best described by the cue abstraction model; likewise, 78.2% of the participants in the justification condition were best described by cue abstraction. Did the chosen strategy also influence how accurately participants judged the test items?

On average, participants classified to the cue abstraction model did not make more accurate judgments in test, $F_{\text{Strategy}}(1, 103) = 1.77$, $p = 0.186$, $\eta^2 < 0.01$, $\text{BF}_{\text{Strategy,0}} = 0.495$, nor were participants more accurate in the justification than in the control group, $F_{\text{Cond}}(1, 103) = 1.3$, $p = 0.256$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond,0}} = 0.525$. Test block neither influenced how accurate participants' judgments were, $F_{\text{Block}}(2.64, 272) = 1.88$, $p = 0.141$, $\eta^2 < 0.01$, $\text{BF}_{\text{Block,0}} = 0.363$. Yet, an interaction between judgment strategy and condition suggested that in the control group, participants classified to the cue abstraction model made more accurate judgments than participants classified to the exemplar model, whereas in justification, participants classified to the exemplar model were more accurate than participants classified to the cue abstraction model, $F_{\text{Strategy x Cond}}(1, 103) = 7.37$, $p = 0.008$, $\eta^2 < 0.05$. However, a corresponding BF analysis provided little support for this finding, $\text{BF}_{\text{Strategy x Cond,0}} = 1.172$.

Descriptively, participants made more consistent judgments if they were assigned to a cue abstraction strategy than if they were were assigned to an exemplar model. Similar to Experiment 1, an ANOVA indicated that participants who justified their judgments were not more consistent than participants in the control group, $F_{\text{Cond}}(1, 103) = 0.58$, $p = 0.447$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond,0}} = 0.205$. Which strategy participants followed neither influenced how consistently they pursued the strategy, $F_{\text{Strategy}}(1, 103) = 2.95$, $p = 0.089$, $\eta^2 < 0.03$, $\text{BF}_{\text{Strategy,0}} = 0.74$. Finally, justification did not interact with strategy use, that is participants classified to the cue abstraction model did not make more consistent judgments in justification than in the control group, nor did participants classified to the exemplar model make less consistent judgments in justification than in the control group, $F_{\text{Strategy x Cond}}(1, 103) = 0.99$, $p = 0.321$, $\eta^2 < 0.01$, $\text{BF}_{\text{Strategy x Cond,0}} = 0.07$. In sum, ...

Analyzing confidence in one's judgment. As a manipulation check, we further analyzed to what extent participant's confidence dropped in justification trials compared to the trials preceding or following the justification using the same method as in Experiment 1. Overall, participants were not less confident in the justification condition than in the control condition, $F_{\text{Cond}}(1, 108) = 3.42$, $p = 0.067$, $\eta^2 < 0.03$, $\text{BF}_{\text{Cond,0}} = 1.185$, but participnats were less confident in justification, $F_{\text{Trial}}(1.33, 143.31) = 9.18$, $p = 0.001$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial,0}} = 68.6$. Specifically, confidence decreased more strongly on justification trials for participants who had to justify their judgment than for participants in the control group, $F_{\text{Trial x Cond}}(1.33, 143.31) = 8.79$, $p = 0.001$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial x Cond,Trial}} = 137.62$. As in Experiment 1, we also put equality constraints on the factor trial type, separately for each condition. This analysis similarly preferred the unconstrained model for the justification condition (BF < 0), but rejected the unconstrained model in the control condition (BF = 27.8).

On average, participants were less accurate in justification than in the control group, $F_{\text{Cond}}(1, 108) = 2.09$, $p = 0.151$, $\eta^2 < 0.02$, but a corresponding Bayesian ANOVA provided less support for this hypothesis, $\text{BF}_{\text{Cond,0}} = 0.685$. However, participants did not make more accurate judgments in trials with a justification than in trials preceding or following a justification, $F_{\text{Cond}}(1.85, 199.71) = 2.38$, $p = 0.099$, $\eta^2 < 0.01$, $\text{BF}_{\text{Trial,0}} = 0.3$. Finally, accuracy did not depend on trial type and trials following a justification did not lead to more accurate judgments in justificaiton than in the control group, $F_{\text{Cond}}(1.85, 199.71) = 1.77$, $p = 0.175$, $\eta^2 < 0.01$, $\text{BF}_{\text{Cond x Trial,0}} = 0.193$. Setting equality constraints rejected the unrestricted model for the control group (BF = 1.9) and

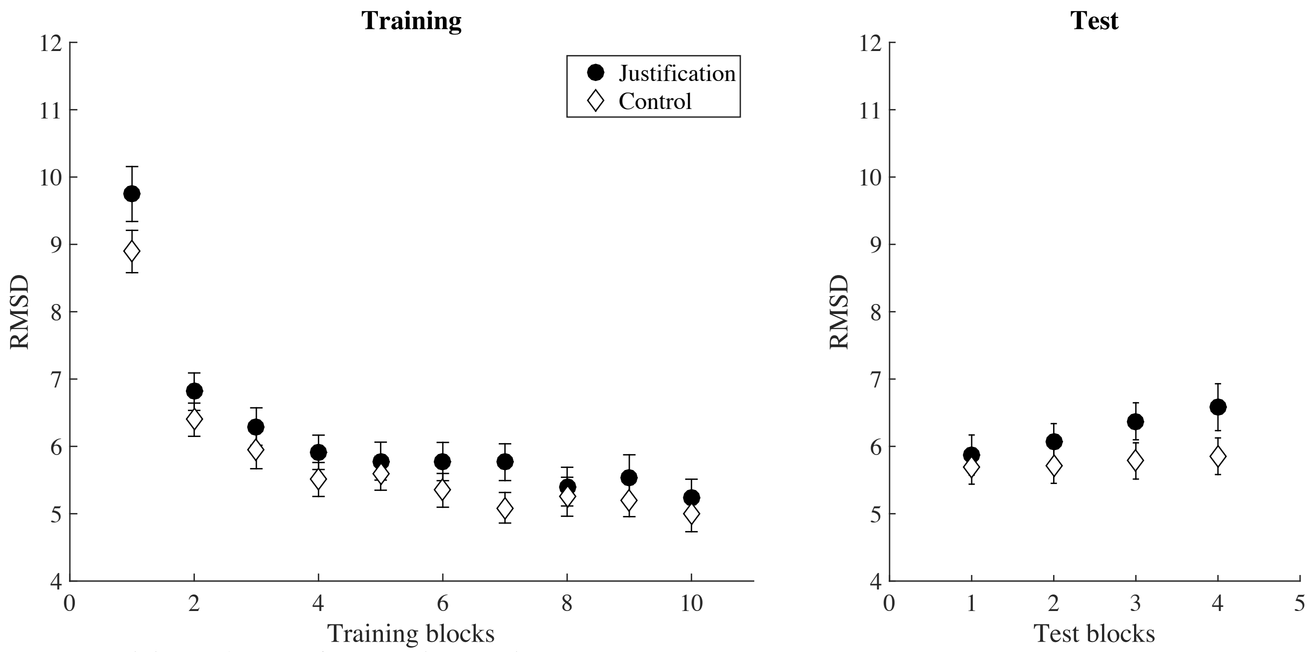

Figure 3. Training and test performance in Experiment 2.

Table 4
Performance and strategy consistency by strategy (cue abstraction or exemplar) in Experiment 1 (Multiplicative Environment) and Experiment 2 (Linear Environment). Standard Deviations in Parentheses.

	Experiment 1			Experiment 2	
	Justification	Verbalization	Control	Justification	Control
Strategies					
Guessing	1	0	0	2	1
Cue abstraction	24	17	21	43	48
Exemplar	24	30	27	10	6
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 the judgment tasks was measured as the Root Mean Squared Deviation.

did not provide sufficient evidence for the hypothesis that trial type influences accuracy in justification (BF = 1.9995).

General Discussion

Providing satisficing reasons for the decision taken is a common duty in professional and private life. Such justifications may provide valuable feedback for others or give important insight into the judgment process and as such have been implemented as tools to improve judgment quality. Yet, our major results indicate that asking for a justification may have less pronounced effects on the decision process and judgment quality as previously suggested. Specifically, we sought to understand how process accountability affects the cognitive processes underlying human judgments and why it may sometimes hurt or benefit perforemance.

In two experiments, we asked our participants to justify their judgments after a random sample of learning trials. In the first experiment, we expected that justifying the judgment process may encourage a higher reliance on cue abstraction instead of relying on exemplar memory. This shift may in turn harm performance in a configural judgment task that is better solved by exempalr memory. Yet, our results suggest that participants who had to justify their judgment neither made less accurate judgments compared to a verbalization and a control condition without justification, nor did judgment strategies change as a function of jusftification. In a second experiment, we further investigated to what degree process accountability may prove benefitial in an elemental judgment task in which a cue abstraction strategy leads to a better performance. In line with the results from the first experiment, participants who had to justify the judgment process did not make more accurate judgments compared to a control group without justification. Descriptively, the results point more towards the opposite view that process accountability led to less accurate judgments. Finally, process accountablity did not encourage the more consistent use of a cue abstraction strategy.

Overall, these results do not match well with findings from previous studies that predominantly find beneficial effects of process accountability (XX;XX, but see XX). One reason for this divergence is potentially that previous studies focussed mostly on the distinct effects of process compared to outcome accountability, and did not include a control condition without any accountability instruction. In our study, we were particularly interested in distinguishing the benefit of process accountability from a judgment process in which participants do not have to justify their judgment and failed to find an impact of process acountability. Matching our findings Siegel-Jacobs and Yates (1996) found that process accountability may not show strong benefitial effects compared to no accountability. Instead, ther results suggest that it is the negative impact of outcome accountability that worsens judgment performance.

One reason for why process accountability did not alter judgment performance is possibly that holding participants accountable for the judgment process did not strongly motivate participants to change their judgment strategy towards a cue abstraction strategy. In both experiments, we did not find that participants justifying their judgment process were more likely to adopt a cue abstraction strategy, nor were their judgments more consistent. Yet, in both experiments process accountability influenced how confident participants were in their judgments, although this effect did not carry on to a lower judgment accuracy in the next trial. These results may indicate that only having to justify a limited number of trials during training is not enough to change the judgment policy but just makes people doubt in that trial. Possibly, people come up with a justification after their judgment without changing the judgment process (Lerner & Tetlock, 1999). This resonates with research showing that people sometimes lack insight into their own judgment processes and give instead plausible responses (Haidt, 2001; Nisbett & Wilson, 1977).

Although we did not find an effect of process accountabilty on judgment accuracy, it is possible that different implementations of process accountability may produce a stronger effect on judgment accuracy. Past research has manipulated process accountability in a variety of ways ranging from announcing a possible report later to an announced interview at the end of the task to videotaping the judgment process (XX; XX). Those manipulations vary in the frequency of expected justifications, the degree of social pressure involved, and whether the justification occurs before or after the decision. In our study, we asked participants to repeatedly justify the judgment after they made a decision and we induced social pressure by explaining that the justifications will be reviewed by another person. Potentially, expecting an interview with another person more strongly increases social pressure than expecting another person to read one's justifications. Future research may investigate more systematically which factors make people reliably feel accountable for the decision process, contrasting process accountability as well with a condition without accountability instructions.

Taken together, our experiments provide little support for the commmon idea that providing a satisfying explanation towards others improves decision quality because people more systematically weigh and integrate all information. Instead providing a satisfying explanation may only make people feel more insecure about their decision.

References

Ashton, R. H. (1990). Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification. *Journal of Accounting Research*, *28*(3), 148. doi:10.2307/2491253

Ashton, R. H. (1992). Effects of justification and a mechanical aid on judgment performance. *Organizational Behavior and Human Decision Processes*, *52*(2), 292–306. doi:10.1016/0749-5978(92)90040-E

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. doi:10.1037/0022-3514.90.6.927

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. doi:10.1037/a0023466

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.

Hagafors, R. & Brehmer, B. (1983). Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Perfor-*

mance, *31*(2), 223–232. doi:10.1016/0030-5073(83)90122-8

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. doi:10.1037//0033-295X.108.4.814

Halberstadt, J. B. & Levine, G. M. (1999). Effects of Reasons Analysis on the Accuracy of Predicting Basketball Games. *Journal of Applied Social Psychology*, *29*, 517–530. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.1999.tb01399.x/abstract

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. doi:10.1177/0956797612463581

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. doi:10.1037/a0037989

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. doi:10.1037/xlm0000241

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*(1), 259–298. doi:10.1016/j.cognition.2007.02.003

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. doi:10.1207/s15327663jcp0404_01

Lagnado, D. a., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–83. doi:10.1037/0096-3445.135.2.162

Langhe, B. D., van Osselaer, S. M., & 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. doi:10.1016/j.obhdp.2011.02.003

Lerner, J. S. & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. doi:10.1037/0033-2909.125.2.255

Levine, G. M., Halberstadt, J. B., & Goldstone, R. L. (1996). Reasoning and the Weighting of Attributes in Attitude Judgments. *Journal of Personality and Social Psychology*, *70*(2), 230–240.

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. doi:10.1037/a0026084

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. doi:10.3758/s13415-013-0215-1

Mero, N. P., Motowidlow, S. J., & Anna, A. L. (2003). Effects of accountability on rating behavior and rater accuracy. *Journal of applied social psychology*, *33*(12), 2493–2514. doi:10.1111/j.1559-1816.2003.tb02777.x

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). 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. doi:10.1037/0033-295X.84.3.231

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), 1371–1384. doi:10.1037/0278-7393.32.6.1371

Pachur, T. & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. doi:10.1016/j.cogpsych.2012.03.003

Reilly, B. A. & Doherty, M. E. (1989). A note on the assessment of self-insight in judgment research. *Organizational Behavior and Human Decision Processes*, *44*(1), 123–131. doi:10.1016/0749-5978(89)90038-1

Reilly, B. A. & Doherty, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes*, *53*(3), 285–309. doi:10.1016/0749-5978(92)90067-H

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, *16*(8), 989–997. doi:10.1002/acp.930

Schooler, J. W. & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71. doi:10.1016/0010-0285(90)90003-M

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. doi:10.1006/obhd.1996.0001

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. doi:10.1037/0022-3514.52.4.700

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. doi:10.1037/a0016690

von Helversen, B. & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*(1), 73–96. doi:10.1037/0096-3445.137.1.73

Weldon, E. & Gargano, G. M. (1988). Cognitive Loafing: The Effects of Accountability and Shared Responsibility on Cognitive Effort. *Personality and Social Psychology Bulletin*, *14*(1), 159–171. doi:10 . 1177 / 0146167288141016. arXiv: 0803973233

Wilson, T. D., Hodges, S. D., & LaFleur, S. J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology*, *69*(1), 16–28. doi:10.1037/0022-3514.69.1.16

Wilson, T. D. & Kraft, D. (1993). Why Do I Love Thee?: Effects of Repeated Introspections about a Dating Relationship on Attitudes toward the relationship. *Personality and Social Psychology Bulletin*, *19*(4), 409–418. doi:10.1177/0146167293194006

Wilson, T. D. & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192. doi:10.1037/0022-3514.60.2.181