Received: 29 May 2019 | Revised: 22 December 2019 | Accepted: 22 December 2019

DOI: 10.1002/bdm.2173



WILEY

RESEARCH ARTICLE

# Dyadic judgments based on conflicting samples: The failure to ignore invalid input

Klaus Fiedler[1] | Tobias Krüger[2] | Alex Koch[3] | Florian Kutzner[1]

[1]Department of Psychology, University of Heidelberg, Heidelberg, Germany

[2]Department of Psychology, Neu-Ulm University of Applied Sciences, Hochschule Neu-Ulm, Neu-Ulm, Germany

[3]Department of Psychology, University of Chicago Booth School of Business, Chicago, IL

**Correspondence**
Klaus Fiedler, Department of Psychology, University of Heidelberg, Heidelberg, Germany.
Email: kf@psychologie.uni-heidelberg.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: Fi 294/26-1, Fi 294/23-1

## Abstract

The present research demonstrates a so far unrecognized impediment of group performance, metacognitive myopia (Fiedler, 2012). Judges and decision-makers follow the given samples of information uncritically and neglect the metacognitive assessment of the samples' validity. Applying this notion to dyadic judgments, we instructed dyads to jointly estimate conditional probabilities *p* (Win|A) and *p* (Win|B) of Lotteries A and B. One person per dyad experienced a valid sample (winning rates conditional on lotteries). The other person experienced an invalid, reverse sample (lotteries conditional on winning). Whereas valid samples provide unbiased estimates of lotteries' winning probabilities, invalid samples can greatly misrepresent the association of winning and lotteries (depending on lottery base rates). Across three experiments, metacognitive myopia—both at the individual and at the dyadic group level—prevented participants from discriminating valid and invalid samples. Group judgments were biased toward erroneous implications of invalid samples, reflecting an equality bias among unequal group members.

**KEYWORDS**

collective reasoning, conditional sampling, epistemic vigilance, metacognitive myopia, sampling bias

## 1 | INTRODUCTION

Democratic societies rely heavily on the validity of group judgments and decisions. However, groups can only outperform individuals when group members communicate effectively. However, in aggregating socially distributed knowledge, they must be sensitive to the validity and the novelty of communicated information, integrating valid and new information but discarding information that is demonstrably invalid (Bonner & Baumann, 2012).

A long research tradition on group decision-making has shown indeed that groups outperform individuals on tasks with a demonstrably valid solution; the correctness of which can be recognized at least by a subset of (at least two) group members (Laughlin & Ellis, 1986). However, small group research has also provided evidence for the conspicuous failure to exploit a group advantage (Kerr & Tindale, 2004) even though a validity criterion is comprehensible by all intellectual standards but unlikely to be jointly met by two or more people. For instance, groups have been shown to perform poorly on hidden-profile tasks (Schulz-Hardt & Mojzisch, 2012) as they fail to jointly assess the full knowledge that is unevenly distributed over group members. They disregard the value of unshared knowledge held by single group members or minorities and thereby fail to jointly discern a demonstrable rule required to solve hidden-profile tasks. In a similar vein, research on advice taking (Yaniv, Choshen-Hillel, & Milyavsky, 2009) reflects a preference for advice from people who share one's own standpoint or sources, although advice takers ought to

understand that independent advice is logically more informative. Social motives (i.e., social approval and mutual respect; Wittenbaum & Park, 2001) may intensify the difficulty of joint rule extraction.

Although shared-information bias and unwarranted trust in redundant advice have been the focus of expanded prior research, (Larson, Foster-Fishman, & Keys, 1994; Yaniv et al., 2009), the present approach addresses a long overlooked deficit called "metacognitive myopia" (MM). A growing body of evidence points to a conspicuous neglect of metacognitive monitoring and control functions (Ackerman & Thompson, 2017; Nelson, 1996). This deficit in quality control of one's own cognitive processes constitutes a serious impediment of rational thought at the individual level (Fiedler, 2000, 2012; Fiedler, Hütter, Schott, & Kutzner, 2019, Fiedler, Hofferbert, & Wöllert, 2018, Fiedler, Hütter, et al., 2019; Unkelbach, Fiedler, & Freytag, 2007). The experiments reported below extend MM to dyadic groups, testing whether dyads fed with conflicting information can jointly overcome MM when they are sensitized to validity issues.

Although judges and decision-makers can be remarkably accurate in utilizing even complex samples of given information, MM renders them uncritical and naïve regarding the validity of the sampled information (Fiedler, 2012). Even when information is obviously biased, it is nevertheless incorporated for uncritical inferences (Fiedler, 2012; Fiedler, Brinkmann, Betsch & Wild, 2000; Unkelbach et al., 2007). When different information sources vary in validity, an "equality bias" (Mahmoodi et al., 2015) may prevent people from separating the wheat (valid information) from the chaff (invalid information).

For example, risk estimates were seriously biased toward obviously invalid advice, despite the presence of contrasting valid advice and despite explicit debriefings and warnings not to fall prey to invalid advice (Fiedler, Hütter, et al., 2019). Participants' estimates of the likelihood of breast cancer given a positive mammogram were strongly influenced by a highly biased sample, which included patients with breast cancer at a much higher base rate (50%) than in the population (4%). This was true even when unbiased samples were available and when participants claimed to understand that the breast cancer base rate in the population was highly relevant (Fiedler, Brinkmann, Betsch & Wild, 2000). A long list of similar findings from diverse paradigms testifies to people's conspicuous failure to exclude information that a metacognitive check ought to disclose as invalid (Fiedler, 2000; Juslin, Winman, & Hansson, 2007; Stewart, Chater, & Brown, 2006).

Would the social context of groups or collective settings afford a remedy to the MM deficit? The literatures on epistemic vigilance (Sperber et al., 2010) and on evolutionary origins of rational reasoning (Cosmides & Tooby, 2013) suggest that social settings can sensitize people for misleading and invalid information and thus trigger more critical validity checks. However, a social setting may not be enough (Kerr & Tindale, 2004). According to a long-established research program by Laughlin and colleagues (Laughlin & Adamopoulos, 1980; Laughlin, Carey, & Kerr, 2008; Laughlin & Ellis, 1986; Laughlin, Hatch, Silver, & Boh, 2006; see also Bonner & Cadman, 2014), groups outperform individuals only when at least two group members understand a demonstrably correct solution. Whether this criterion is met depends on the degree to which groups facilitate (a) individual-level inferences and (b) group-level communication of the correct solution. Thus, from Laughlin's demonstrability perspective, overcoming MM depends on the strength of these two facilitation effects. More generally, whether a group advantage is borne out or not constitutes an open empirical question; MM may persist at group level when the twofold facilitation effect is insufficient.

For a suitable experimental test, we created a sample-based decision task that prior research had shown to give rise to distinct MM effects. Both members of a dyad actively sample observations about the outcomes of two lotteries, drawing from the same universe but from different perspectives. Whereas one person samples outcomes (Win vs. Fail) conditional on lotteries (A vs. B), the other person samples lotteries (A vs. B) conditional on outcomes (wins vs. nonwins). Note that the former sampling procedure yields proportions $P$ (Win|A) and $P$ (Win|B) that afford valid, unbiased estimates of the true winning probabilities $p$ (Win|A) and $p$ (Win|B). In contrast, the inverse proportions $P$(A|Win) and $P$(A|Fail) obtained by the other, invalid sampler are biased toward $p$(A) base rates. We manipulate the base rates $p$(A) and $p$(B) such that the inverse sample proportion $P$(A|Win) obtained by invalid samplers can diverge markedly from the valid sampler's unbiased proportion $P$ (Win|A).

After the sampling stage, both dyad members first provide their individual estimates $p^*$(Win|A) and $p^*$(Win|B) of the lotteries' winning rates before they finally provide group estimates and choose one lottery they jointly prefer to play. The apparent conflict of a valid and an invalid sampler should trigger epistemic vigilance and sensitize dyads to validity problem. Dyads ought to base their judgments and choices on the valid sample and discard the invalid sample. Yet we suspect that MM may not allow dyads to follow a validity-driven strategy. Rather, MM may carry over from the individual to the group level, misleading dyads to base their joint judgments on a compromise that contaminates valid with invalid samples.

Previous research on individual-level MM (Fiedler, 2012; Fiedler, Schott, et al., 2019) suggests that dyads may not discard invalid samples but naïvely utilize all sampled information, giving substantial weight to both opinions (Mahmoodi et al., 2015). Invalid samplers' estimates may serve as numerical anchors for collective estimates. Group judgments may thus reflect an uncritical opinion-negotiation process (Fiedler et al., 2018; Mahmoodi et al., 2015).[1]

One might question participants' ability to understand the logic of conditional sampling. The depicted research may thus speak to mundane cognitive competence rather than a metacognitive deficit. The dyads' failure to follow valid samplers and discard invalid samplers may simply reflect their common inability to distinguish between $p$ (Win|A) and $p$(A|Win). This objection is not applicable to our theoretical approach for two reasons.

First, metacognition is not fundamentally different from cognition. We rather define metacognition as those self-critical cognitive processes that serve to monitor and control the quality of one's own mental operations (Ackerman & Thompson, 2017; Nelson, 1996).

[1]In addition to numerical anchoring, social motives may be involved, such as a superficial fairness rule or the motive to get along with each other. Failure to control for such motives may contribute to MM.

Critical assessment of information invalidity is thus by definition metacognitive, because it serves a monitoring and control function. Even when it turns out that (some) participants do not understand conditional probabilities, this remains a metacognitive deficit, for the logic of conditional probabilities is essential for validity monitoring and control.

However, second, conditional probabilities are involved in countless everyday tasks; they do not exceed common levels of intelligence. Through multiple everyday experience, people can easily understand that the probability *p* (right-handed|female) that a female person is right-handed is >80%, because the *p* (right-handed) base-rate is so high, even though the reverse conditional *p* (female|right-handed) that right-handed persons are female is close to 50%. Every child understands that all dogs have tails although tails need not belong to dogs. Participants in a conditioning experiment, even animals, have no problem to learn that conditioned stimulus is always followed by unconditioned stimulus although unconditioned stimulus may often be preceded by other stimuli. Although all these conditional probability assessment tasks have a uniquely demonstrable correct solution (Laughlin, 2011; Laughlin & Ellis, 1986), MM may prevent even intelligent people from thinking and talking about something as simple as the asymmetry of conditional probabilities.

MM need not mean that people explicitly understand that a logical rule applies and nevertheless do not use it.[2] MM first of all means that for some reason, people do not think of and do not systematically utilize rules that belong to their repertoire. For an analogy, the seminal demonstrations of heuristics and biases (Tversky & Kahneman, 1974) never depended on a proof that participants explicitly understand the base rate fallacy, the law of large numbers, or the conjunction fallacies. The evidence rather highlights the fact that most participants do not apply these principles, although the human mind can easily understand these principles. Going beyond the heuristics-and-biases program, MM does not stop with the fact that anchoring biases, conjunction fallacies, or base rate neglect arise in the first place. Rather, MM assumes that a comprehensive explanation of irrational behavior must also explain why such primary mistakes are not discovered and corrected at the metacognitive level, even when individuals are sensitized to the violation of logical rules.

## 2 | AIMS AND PREVIEW OF EXPERIMENTS

Three experiments examined whether individual-level MM in conditional reasoning can be overcome at group level when dyads are fed with conflicting samples of observations about the outcomes (Win vs. Fail) of two Lotteries A and B. One member of each dyad saw a valid sample of both lotteries' (unbiased) winning rates *p* (Win|A) and *p* (Win|B); the other member experienced an invalid sample of information about reverse conditionals, *p*(A|Win) and *p*(B|Win), which are biased toward the lotteries' base rates and therefore logically inappropriate. Individual-level MM would be evident in invalid samplers' tendency to mistake high (vs. low) *p*(A|Win) for high (vs. low) winning rate *p* (Win|A); the influence of invalid samplers on dyadic estimations and choices affords a measure of MM at the group level. The strength of the discrepancy between *p* (Win|A) and *p*(A|Win) was manipulated between experimental groups. The persistence of MM in dyads would be at odds with the notion of epistemic vigilance, which implies that dyads fed with conflicting samples should be sensitized for a possible validity problem and that their discussions should thus find out that only valid samples should be utilized, whereas invalid samples should be discarded.

In all experiments, the superiority of Lottery A over B was set to $\Delta = p\,(\text{Win}|\text{A}) - p\,(\text{Win}|\text{B}) \sim .20$, respectively. In Experiment 1, invalid samplers could only sample observations about winning outcomes; nonwinning outcomes were fully excluded to highlight the logical impossibility to estimate *p* (Win|A) and *p* (Win|B) from such incomplete data. This feature was modified in Experiment 2 (using a larger sample of dyads) and in Experiment 3 (replacing lotteries by another choice task), in which invalid samplers exposed to both winning and nonwinning trials had a fair chance to infer *p* (Win|A) and *p* (Win|B) from the reverse conditionals. We found convergent evidence for MM both at the individual and the group level. As invalid samplers' *p*(A|Win) conditionals varied more strongly than valid samplers' *p* (Win|A) and *p* (Win|B) conditionals; the irrelevant *P*(A|Win) proportion became the strongest predictor of the dyads' final estimates.

## 3 | EXPERIMENT 1

For dyads to overcome MM and to behave rationally, they had to find out through discussion that one sample was invalid for the joint judgment task. If they failed to solve this communicative task, their judgment should represent an unwarranted compromise of both samples. Because only the misleading "reverse" conditionals varied between conditions whereas the appropriate "forward" conditionals were constant; the joint estimates should depend on the invalid samplers' variable experience more than on valid samplers' constant experience.

### 3.1 | Methods

#### 3.1.1 | Participants and design

Fifty-four Heidelberg University students participated for course credit or for payment. In total, 27 pairs were randomly allocated to three conditions. Within dyads, participants were randomly assigned the valid and invalid sampler role.

#### 3.1.2 | Materials and procedure

The basic version of the sampling task was constructed to render the choice task quite easy for valid samplers, setting the lottery winning

[2]Although this may occur occasionally, participants who do not utilize a rule are also unlikely to be momentarily aware of that rule, even when their cognitive capacity enables them to fully understand it.

**TABLE 1** Manipulated parameters, mean contingency judgments (*standard deviations*), and mean WIO for Experiment 2

| | Experiment 1 | | | Experiment 2 | |
|---|---|---|---|---|---|
| **Condition** | **High** | **Med** | **Low** | **High** | **Low** |
| *p* (Win\|A) | .60 | .60 | .60 | .60 | .60 |
| *p* (Win\|B) | .40 | .40 | .40 | .40 | .40 |
| *p*(A\|Win) | .75 | .41 | .27 | .88 | .27 |
| *p*(B\|Win) | .25 | .60 | .73 | .12 | .73 |
| *p**(Win\|A) | .603 (*.172*) | .540 (*.232*) | .528 (*.256*) | .462 (*.170*) | .477 (*.164*) |
| *p**(Win\|B) | .261 (*.071*) | .460 (*.190*) | .354 (*.129*) | .263 (*.143*) | .414 (*.186*) |
| Δ* | +.342 (*.177*) | +.080 (*.223*) | +.173 (*.272*) | +.199 (*.212*) | +.063 (*.197*) |
| WIO | Not available | | | .398 (*.384*) | .479 (*.289*) |

*Note.* Letters A and B refer to lotteries. An asterisk denotes subjective estimates. Note that Δ* = *p**(Win|A) − *p**(Win|B). WIO is an index measuring the weight given to an invalid opinion.
Abbreviation: WIO, weight of invalid opinion.

rates to *p* (Win|A) = .60 and *p* (Win|B) = .40 and, hence, the superiority of A over B to Δ = .60 − .40 = .20 (see Table 1). For invalid samplers in three conditions, the reverse conditional probability varied from *p*(A|Win) = .75 (high) to.41 (medium) to.27 (low). MM should be evident in invalid samplers' estimates and dyadic estimates increasing from low to medium to high.

At the beginning of the experiment, both dyad members were seated in front of separate computers. They read that their task was about "identifying the better one of two Lotteries A and B." Valid samplers saw 50 observations from both lotteries, which prepared them for the question: "When someone played Lottery A or B, did that person win or lose?" In contrast, invalid samplers observed 50 trial outcomes (i.e., winning or losing) along with the question "When someone won or lost, did that person play Lottery A or B?" Participants read that they would judge lotteries first individually and then again as a group, with an extra reward of €2.50 for a correct choice.

*Individual sampling phase*

Next, at the computer, participants saw the sample of 50 observations according to their condition. For valid samplers, each trial started with a black screen. After 1,000 ms, the "Next trial" was announced in the upper middle of the screen, and after another 1,500 ms, the text "This time it was played" appeared in the same position, and 1,500 ms later, the label "Lottery A" or "Lottery B" was presented underneath in the upper part of a blue or orange rectangular frame, respectively. In the lower part of the frame, a row of symbols was running at a high speed in a horizontal field (at a rate of 10 ms/symbol), reminiscent of a real gambling machine. Then the text "The result was" inserted below the frame, and finally, after 1,000 ms, the running symbols were replaced by the feedback "WON" or "NOTHING" (in German "GEWONNEN" vs. "NICHTS") presented for 2 s. The frame would then shrink in four steps and disappear until the start of the next trial.

For invalid samplers, the announcement "Next trial" was followed by "The result was" on the bottom with running symbols and outcome information (winning or not) presented in a black frame. If the choice was not winning, the phrase "Will not be considered" appeared and no further feedback was provided. If the outcome was winning, the blank frame turned into a blue or orange frame labeled "Lottery A" or "Lottery B," respectively. This feature served to highlight the logical unsuitability of invalid samples; it was changed in later experiments.

At the end of the sampling phase, still in computer dialog, both individuals indicated "Which lottery would you prefer?" in a dichotomous choice and, subsequently, provided separate estimates of the lotteries' winning probabilities *p* (Win|A) and *p* (Win|B) in percent, as well as confidence ratings about their answers (endorsement of the statement "My decision is certainly correct" on a 7-point scale). Unfortunately, due to programming error, these individual data were lost in Experiment 1. They will be available in all following experiments.

*Group discussion phase*

Both dyad members were seated at a table in the center of the room and were asked to discuss their sampling experiences with the lotteries. They were reminded of the goal to jointly reach a solution regarding which lottery yielded a higher winning probability and that they would receive €2.50 as a reward for answering correctly. Then they together completed a group questionnaire, including the same questions they had answered individually.[3] Finally, they were thanked, debriefed, and paid.

## 3.2 | Results and discussion

For an overall measure of the influence of the invalid sample on the group decision, we calculated the correlation between the

[3]They also filled out an exploratory questionnaire that can be found in the supplements. However, because this questionnaire mainly dealt with qualitative questions about the experience of the discussion, it did not provide any useful evidence and it was replaced by a more pointed questionnaire in Experiment 2 below.
[4]One-sided tests were conducted when a reverse hypothesis did not make any sense, theoretically.

manipulated *p* (win|A) in the invalid sample (.75, .41, or.27) and the differential group estimate $\Delta^* = p^*(\text{Win}|\text{A}) - p^*(\text{Win}|\text{B})$. A positive correlation supports the hypothesis that joint $\Delta^*$ estimates are biased toward invalid samplers' experienced association *p*(A|Win) between A and winning. The obtained correlation $r$ (df = 25) = +.366, $p_{\text{one-sided}}$ = .030,[4] is consistent with the notion of MM. Invalid samplers' reverse conditionals exerted a significant influence on the group-level lottery estimates, despite their logical unsuitability: The higher the invalid probability *p*(A|win), the higher were the group estimates $\Delta^*$ of the winning probability of Lottery A compared with Lottery B.

Closer analyses revealed that joint estimates of the superiority of A over B were higher in the high *p*(A|Win) condition, $M_{\Delta^*}$ = +0.342, *SD* = 0.177, than in the medium condition, $M_{\Delta^*}$ = +0.080, *SD* = 0.223, and in the low *p*(A|Win) condition, $M_{\Delta^*}$ = +0.173, *SD* = 0.272. Due to the very small samples of Experiment 1, we refrain here from pairwise significance tests between three too small subgroups. Suffice it to mention that across all 27 dyads, the final contingency estimates amounted on average to $M_{\Delta^*}$ = +0.199, *SD* = 0.245, which is significantly above 0, $t(26) = 4.217$, $d = 1.623$, $p < .001$, testifying to the participants' high accuracy motivation and their sensitivity to the sign and size of the true contingency $\Delta$ = +.20.

In summary, Experiment 1 corroborates the original demonstration of MM in conditional inferences (Fiedler, 2008), extrapolating the phenomenon from individual to dyadic reasoning. Although epistemic vigilance should have sensitized judges for the discrepancy between valid and invalid samples, the participants were unable to identify the inapplicability of the invalid sample in their group discussion. As a consequence, their joint estimates were systematically biased toward invalid samplers' opinions, reflecting an uncritical compromise of valid and invalid information.

## 4 | EXPERIMENT 2

In Experiment 2, an improved design allowed for a more systematic test of the theoretical ideas, for which Experiments 1 already provided initial support. First, to increase statistical power, we increased the number of dyads to 57, of which 30 and 27 were assigned to high versus low *p*(A|Win), respectively. Second, a questionnaire was constructed to assess epistemic vigilance in a reasonable response format (see Appendix). Third, we exposed both valid and invalid samplers to all 2 × 2 combinations of lotteries (A vs. B) and outcomes (Win vs. Not Win), rather than restricting invalid samples to winning outcomes as in Experiment 1. Although well-motivated in Experiments 1, this may have undermined invalid samplers' motivation to solve the contingency problem. In Experiment 2, all participants could observe the full 2 × 2 distribution lotteries and outcomes.

Most importantly, this modified task structure enabled a new distinction of MM at the level of individual participants (in the invalid sampling condition) and MM at the group level. Given a random sample of all four joint frequencies of lotteries and outcomes, there are two simple and straightforward ways of arriving at unbiased $\Delta$ estimates and avoiding the anomaly demonstrated above. First, to evade the biasing influence of invalid samples, dyads could simply decide to follow the valid sample and to fully ignore the invalid sample. The failure to reach this insight in dyadic discussion reflects MM at the group level.

However, second, there is also a fair individual-level chance for invalid samplers to avoid the mistake in conditional reasoning before dyadic interaction comes into play. Although the conditional direction of stimulus pairs is reversed for invalid samplers—they are exposed to lottery labels conditional on outcomes rather than outcomes conditional on lotteries—it will soon be apparent that the resulting 2 × 2 joint frequencies allow for an unbiased estimation of $\Delta = p$ (Win|A) − $p$ (Win|B) in all experimental conditions (see manipulation check below). Thus, overcoming MM at the individual level merely calls for estimating $\Delta$ from the available joint frequencies.

Yet given that valid and invalid samplers are exposed to unbiased samples of joint frequencies, why should invalid samplers arrive at biased inferences in the first place? A possible answer is apparently that the underlying cognitive process does not rely on joint frequencies but on encoded quantities that vary between valid and invalid samplers. Apparently, valid samplers' estimates reflect continuous updating (Hogarth & Einhorn, 1992) of conditional winning expectancies for A versus B, whereas invalid samplers' estimates reflect continuous updating of another conditional quantity, namely, expectancy of A given winning versus losing outcome. Although based on the same raw information, the two types of encoded conditional expectancies are essentially different. Once the sampling history of these encoded expectancies is forgotten, it seems to be impossible to back-translate one conditional memory code into the other or into joint frequencies.

Thus, when later asked to estimate *p* (Win|A) and *p* (Win|B), valid samplers can rely on their encoded conditional estimate, whereas invalid samplers commit the mistake to follow the stronger association of winning with Lottery A than with Lottery B. An appropriate remedy—and a means of overcoming MM—would be to store joint frequencies that afford an unbiased estimate of $\Delta$. The failure to use this remedy and the continued confusion of two essentially different conditional memory codes suggest a mechanism for individual-level MM.

In any case, MM can be avoided both at individual level and at dyadic level. Both levels can be distinguished as we assess individual estimates prior to the final dyadic estimates.

### 4.1 | Methods

#### 4.1.1 | Participants and design

One-hundred fourteen male and female students of the University of Heidelberg were recruited via a local online recruiting platform. They participated either for course credit or for payment (at a rate of €8/hr). Pairs of consecutively appearing participants were randomly allocated to the valid and the invalid sampling condition. If an odd number of participants appeared, the leftover person performed the sampling and the individual estimation task individually. All pairs ($N$ = 57) or

singular participants ($N$ = 22) were then randomly assigned to the high or low $p$(A|Win) condition (see Table 1). Although the lottery winning rates were always constant and favored Lottery A, $\Delta = p$ (Win|A) − $p$ (Win|B) = .60 − .40 = .20, the reverse conditionals varied between $p$ (A|Win) = .878 and $p$(A|Win) = .268 (see Table 1).

### 4.1.2 | Materials and procedures

The same instructions and computer software were used as in Experiment 1, with one exception: In the invalid sampling condition, nonwinning trials were not truncated, but each trial started with a participant's choosing either a winning or a nonwinning outcome, followed by feedback about what lottery had been played. Both invalid and valid samplers were thus exposed to all four stimulus combinations. Although the joint frequencies varied markedly between conditions, they afforded equally accurate estimates of $\Delta = p$ (Win|A) − $p$ (Win|B) in all conditions, as will be apparent from the manipulation check below. Thus, the advantage of A over B was visible from all viewpoints.

After individual (valid and invalid) samplers had gathered their stimulus observations and provided their individual preferences and estimates on separate computers, they engaged in a dyadic discussion and finally provided their joint paper-and-pencil estimates.

The first question of the postexperimental questionnaire (see Appendix) asked for a multiple-choice estimate of how long the dyad had discussed (<1 min, 1–3 min, 3–5 min, and >5 min). Six of the remaining questions referred to epistemic vigilance solicited by conflicting opinions: To what degree did you finally agree about which lottery is better (not at all/completely)? The samples provided by the two of us initially suggested lotteries that were (very similar/incompatible)? How big was the gap that had to be bridged between the two standpoints (very strong/not existing)? Did you give an equal weight to both opinions, or did you decidedly follow one opinion you recognized to be the correct one (equally weighted/one-sided)? What determined the joint evaluation, the discussants' relative eagerness, or the informative value of the samples (samples/eagerness)? Was the unequal rate of Lottery A/B or of winning/not winning a topic of discussion that influenced your estimate (definitely no/definitely yes)? Only Item 7 did not refer to epistemic vigilance: Whose sample was logically more suitable for the evaluation of lotteries (my sample/the other sample)?

## 4.2 | Results and discussion

### 4.2.1 | Sample statistics

An analysis of the samples drawn by valid and invalid samplers (missing data for one invalid sampler) provides a successful manipulation check. As intended, for invalid samplers the sampled proportion of A lotteries given a winning outcome in the high $p$(A|Win) condition ($M_{P(A|Win)}$ = 0.840, $SD$ = 0.067) greatly exceeded the sampled proportion in the low $p$(A|win) condition ($M_{P(A|Win)}$ = 0.271, $SD$ = 0.091), $t$(55) = 26.772, $d$ = 7.220, $p$ < .001, mirroring the manipulation of $p$(A|Win) = .878 versus.286.

Nevertheless, both valid and invalid samplers were exposed to samples from the same underlying distribution, despite the different perspective. Therefore, observed contingencies $\Delta_{obs} = a/(a + b) - c/(c + d)$, computed from the observed joint frequencies $a$, $b$, $c$, and $d$, approximated the manipulation of $\Delta = .2$ in all conditions ($M_{\Delta obs}$ between 0.14 and 0.18). For valid samplers, the average sampled $\Delta_{obs}$ was $M_{\Delta obs}$ = 0.162, $SD$ = 0.210, $t$(56) = 5.803, $d$ = 1.537, $p$ < .001, with little difference between the high, $M_{\Delta obs}$ = 0.149 ($SD$ = 0.131) and low $p$(A) conditions, $M_{\Delta obs}$ = 0.176 ($SD$ = 0.276), $t$(55) = –0.488, $d$ = –0.130, $p$ = .627. For invalid samplers, the experienced joint frequencies also reflected the true contingency, $M_{\Delta obs}$ = 0.155, $SD$ = 0.140, $t$ (56) = 8.377, $d$ = 2.219, $p$ <\ .001, and this was similarly the case in the high, $M_{\Delta}$ = 0.146 ($SD$ = 0.162), and in the low $p$(A|Win) condition, $M_{\Delta obs}$ = 0.165 ($SD$ = 0.112), $t$(55) = –0.504, $d$ = –0.135, $p$ = .616. Thus, had participants estimated Δ from the observed joint frequencies, they could have evaded MM and provided equally accurate estimates and choices in all conditions.

### 4.2.2 | Individual estimates

Although the samples encountered in all conditions reflect the advantage of A over B to a similar degree, the resulting winning estimates of A and B varied strongly with experimental conditions. Valid samplers' lottery winning estimates, $M_{p*(Win|A)}$ = 0.424, $SD$ = 0.163 and $M_{p*(Win|B)}$ = 0.237, $SD$ = 0.155, and the corresponding Δ* scores clearly reflected the advantage of Lottery A, $M_{\Delta*}$ = 0.187, $SD$ = 0.168, $t$(56) = 8.413, $d$ = 2.229, $p$ < .001.

In contrast, invalid samplers' Δ* estimates, $M_{p*(Win|A)}$ = 0.532, $SD$ = 0.242, $M_{p*(Win|B)}$ = 0.445, $SD$ = 0.231, yielding $M_{\Delta*}$ = 0.087, $SD$ = 0.387, $t$(55) = 1.682, $d$ = 0.449, $p$ = .098, were to a lesser degree determined by the observed contingency of winning rates on lotteries. Because invalid samplers focused on A conditional on winning, their Δ* estimates were strongly affected, $r$ ($df$ = 54) = .475, $p_{one\text{-}sided}$ < .001, by the logically inappropriate $p$(A|Win).

Separate regression analyses were conducted for valid and invalid samplers, using individual Δ* estimates as criterion and two predictors: first, the normatively appropriate sample difference $\Delta_{obs} = a/(a + b) - c/(c + d)$ and second the misleading proportion of Lottery A given that a win had been sampled, $a/(a + c)$. As expected, the appropriate predictor largely determined valid samplers' estimates, $r$ = .659, $\beta$ = .574, $t$(54) = 3.029, $p$ < .001, compared with $r$ = .581, $\beta$ = .101, $t$(54) = 0.533, $p$ = .596, for the other predictor ($df$ = 54 because two valid samplers never sampled Lottery B; thus, $\Delta_{obs}$ could not be computed for these two samplers). In contrast, invalid samplers' Δ* estimates reflect to a much lesser degree the normatively correct cue, $r$ = .253, $\beta$ = .236, $t$(55) = 2.052, $p$ = .045. Their regression weights were much higher for the $a/(a + c)$ predictor, $r$ = .493, $\beta$ = .485, $t$(55) = 4.214, $p$ < .001. Thus, rather than extracting the unbiased contingency inherent in the available joint frequencies, $a$, $b$, $c$, and $d$, they mistook the

conditional rate of Lottery A given a winning outcome for the inverse conditional rate of winning given specific lotteries. This failure to distinguish between reverse conditionals is indicative of a serious failure to overcome MM at the individual level.

### 4.2.3 | Dyadic estimates

Let us again finally consider the dependency of the dyadic estimates on the inverse conditional *p*(A|Win) that was manipulated for invalid samplers. Consistent with Experiment 1, the correlation between $\Delta^* = p^*(\text{Win}|\text{A}) - p^*(\text{Win}|\text{B})$ and the high versus low *p*(A|Win) amounted to $r\ (df = 55) = +.319$, $p_{\text{one-sided}} = .008$.

The general tendency to estimate the winning rate of A higher than the winning rate of B was markedly stronger in the high, $M_{\Delta^*} = 0.199$, $SD = 0.212$, than in the Low *p*(A|Win) condition, $M_{\Delta^*} = 0.063$, $SD = 0.197$, $t(55) = 2.493$, $d = 0.666$, $p = .016$ (see Table 1). Thus, the manipulation of *p*(A|Win) in the invalid sample, which is logically irrelevant to estimating $\Delta = p\ (\text{Win}|\text{A}) - p\ (\text{Win}|\text{B})$, again had a profound effect on dyadic estimates.

The average weight of invalid opinion (WIO) was $M = 0.442$ ($SD = 0.275$), $t(41) = 10.42$, $d = 3.215$, $p < .001$; WIOs above zero testify to MM. Similar WIOs were found for high, $M = 0.423$, $SD = 0.231$, and low *p*(A|Win), $M = 0.460$ ($SD = 0.317$), $t(40) = -0.432$, $d = -0.35$, $p = .668$.[5]

Thus, although valid samplers arrived at correct Δ estimates derived from unbiased observed winning rates of both lotteries, their discussions with invalid samplers resulted in a shift of the final dyadic judgments toward invalid samplers' distortions. As the manipulation of *p*(A|Win) misled invalid samplers to overstate or understate the winning rate of Lottery A (relative to B) in the high and low *p*(A|Win) condition, respectively, this logically irrelevant information exerted a similarly strong impact on the final evaluations as the logically relevant information held by the valid samplers. Moreover, because invalid samplers' evaluations varied more strongly (due to the manipulation) than valid samplers' (generally accurate) evaluations, the invalid samples dominated the final dyadic judgments. In a regression analysis, invalid samplers' individual Δ* estimate was a stronger predictor of the dyadic Δ*, $\beta_{\text{invalid}} = .724$, $r = .614$, $t(55) = 7.072$, $p < .001$, than valid samplers' individual Δ*, $\beta_{\text{valid}} = .363$, $r = .143$, $t(55) = 3.545$, $p = .001$. This means that the second chance to overcome MM at the dyadic level was also missed.

### 4.2.4 | Epistemic vigilance as a remedy to MM

What do responses to the postexperimental questionnaire reveal about discussions leading to this unwarranted compromise? Was there an intuitive understanding of sample validity? Were dyads biased toward invalid samples in spite of the tendency to recognize the logical adequacy of the valid sample? And how about epistemic vigilance? Does the impact of invalid samplers' opinion reflect the insensitivity to sample conflict and the absence of epistemic vigilance?

We collapsed questionnaire items 2, 3, 4, 5, 6, and 8 (see Appendix) into an index of epistemic vigilance ($\alpha = .76$) triggered by conflicts and discrepant sampling input. Item 7, assessing the belief in one's own versus the other sample, was treated as a separate predictor.

To understand the potential influence of epistemic vigilance, we regressed the final dyadic Δ* estimates on three predictors, the *p*(A|Win) manipulation of the invalid sampler, the epistemic vigilance score, and Item 7 (relative belief in own vs. other sample), averaging over responses from both partners in the 57 dyads. Very modest intercorrelations (from −.14 to +.07) showed that predictors were largely independent. All three predictors made a significant contribution to predicting the criterion. Dyadic Δ* estimates of the advantage of A over B were higher when *p*(A|Win) was high rather than low, $r = .319$, $\beta_{\text{condition}} = .252$, $t(54) = 2.148$, $p = .036$, when invalid samples were deemed logically correct, $r = -.355$, $\beta_{\text{logical}} = -.331$, $t(54) = -2.826$, $p = .007$, and, notably, when epistemic vigilance was high, $r = .280$, $\beta_{\text{epist. vigilance}} = .2768$, $t(54) = 2.371$, $p = .021$.[6] Thus, epistemic vigilance did not decrease but, if anything, tended to strengthen the dyads' susceptibility to unwarranted influences by invalid samplers.

## 5 | EXPERIMENT 3

We conducted a last experiment to highlight the robustness and flagrancy of MM in conditional reasoning and to rule out the possibility that participants in preceding experiments were simply undermotivated or overwhelmed by too complex an inference task. The experimental task of Experiment 3 was thus modified and simplified in several respects.

First, we replaced the lottery task by a more socially meaningful contingency problem embedded in a new cover story. We asked participants to figure out whether departments of a larger company that had participated in an educational program were more likely to win a prize than nonparticipating departments. Thus, participants judged *p* (Win Prize|Education) relative to *p* (Win Prize|No Education). Although the lottery task leaves it open whether winning is more likely for Lottery A or B, world knowledge tells us that, if anything, education should increase the chance to win a prize. To the extent that conditional reasoning requires social meaning and causal knowledge as a catalyst, the advantage of valid sampling (from cause to effect) over invalid reasoning (from effect to cause) should become more apparent.

Second, for a test of the impact of causal expectancies, we manipulated the direction of the actual contingency, that is, whether an

[5]The weight of invalid opinion index WIO = [dyadic D*− valid D*] /[invalid D*− valid D*] measures the extent to which the final dyadic estimate of D moves on the way from the valid tothe invalid estimate. WIO scores > 1 (dyadic D*farther away than invalid sampler's D* fromvalid samplers*' D*) and < 0 (dyadic D* deviating from valid samplers' D* in thedirection opposite to invalid samplers' D*) were omitted, in accordance with common practice (Fiedler et al., 2019; Fiedleret al. in press; Hütter & Ache, 2016; Hütter Fiedler, in press). Including these cases led to equivalent results.

[6]Note that *df* = 54 (rather than 55) because one conditional probability by one invalid sampler was missing

educational program served to increase (expected) or decrease (unexpected) the chances to win. Thus, the contingency experienced by valid samplers was either positive, $\Delta = p$ (Win Prize|Education) − $p$ (Win Prize|No Education) = .26 − .05 = +.21, or negative $\Delta = .05 - .26 = -.21$. The reverse contingencies presented to invalid samplers were also positive or negative, but stronger; $p$ (Education|Win Prize) − $p$ (Education|Not Win Prize) was either $\Delta' = .82 - .42 = .40$ or $\Delta' = .18 - .58 = -.40$ (Table 2).

In the valid sampling condition, participants received feedback about the winning rates conditional on whether a selected department had or had not undergone an educational program. Participants in the invalid sampling condition did not receive feedback about the winning conditional on educational program. Instead, they determined how often they wanted to gather evidence on a winning or on a nonwinning department; feedback indicated whether that department had or had not participated in an educational program.

At the end of the sampling stage, both valid and invalid samplers evaluated onscreen the chances of a prize given a department had participated in an educational program. To highlight logical irrelevance, invalid (like valid) samplers were only asked to estimate winning rates given an educational program, although invalid samplers had themselves determined how many winning trials they wanted to consider. When asked to estimate $p$ (Win Prize|Education), they should have protested, and in the dyadic discussion, they should have refrained from using their sample. Metacognitive insight into this fact should have led to low confidence ratings and reluctance of both partners to consider the invalid partner's conditionals.

The complexity of the contingency task was reduced to judging only one conditional, $p$ (Win Prize|Education), which was easy to judge for valid samplers but out of reach for invalid samplers. The other conditional, $p$ (Win Prize|No Education), was not included in the individual judgment task, which consisted of three steps: (a) a dichotomous judgment of whether participating in an education program will lead to a prize or not; (b) a quantitative estimation of $p$ (Win Prize|Education); and (c) a 7-point confidence rating (from 1 to 7). After discussing the joint judgment in the dyad, both partners provided collective estimates of both conditional probabilities, $p$ (Win Prize|Education) and $p$ (Win Prize|No Education).

MM at the individual level should be evident in a distinct interaction between sampling (valid vs. invalid) and (positive vs. negative) contingency (between education and winning the prize) on binary and continuous estimates of $p$ (Win Prize|Education). This interaction pattern reflects that valid samplers' estimates of $p$ (Win Prize|Education) should approximate the valid conditionals of.26 and .05 given positive and negative contingency, respectively. Invalid samplers, in contrast, should provide estimates of $p$ (Win Prize|Education) that are distorted toward their invalid sample experience of $p$ (Education|Win Prize) = .82 and.18 in the positive and negative contingency condition, respectively. Both valid and invalid samplers' naïve trust in whatever sample they observed should be evident in relatively high confidence levels (above the midpoint of the 1–7 scale), with which they estimated $p$ (Win Prize|Education), despite the fact that invalid (but not valid) samplers had themselves determined by sampling how often they observed a department that had won the prize. MM at group level should be evident in dyadic judgments' biased toward invalid samplers' opinions.

## 5.1 | Methods

### 5.1.1 | Participants and design

Eighty-two students from the University of Heidelberg participated either for course credit or for monetary compensation. Consecutive pairs were randomly assigned to the valid and invalid sampler roles, and pairs were randomly assigned to either the positive-Δ or the negative-Δ condition (see Table 2).

### 5.1.2 | Materials and procedure

The procedure was similar to all previous experiments; however, the experimental context was changed from lotteries to departments in a larger company. Both valid and invalid samplers were told that their task was about judging the effectiveness of an educational program. They would be allowed to search the archive of a large company for information about departments winning a prize conditional on their

**TABLE 2** Manipulated parameters and judgment means (*standard deviations*) per condition

| Condition | Positive contingency | Negative contingency | |
|---|---|---|---|
| Manipulated parameters | | | |
| $p$ (Win\|Education) | .26 | .05 | Joint frequencies yielded the same Δ for valid and invalid samplers |
| $p$ (Win\|No Education) | .05 | .26 | |
| $p$ (Education\|Win) | .82 | .18 | |
| $p$ (Education\|Not Win) | .42 | .58 | |
| Mean dyadic judgments (*standard deviations*) per condition (negative contingency inverted) | | | |
| | Positive contingency | Negative contingency | Altogether |
| $p$*(Win\|Education) | .464 (*.165*) | .418 (*.213*) | .450 (*.179*) |
| $p$*(Win\|No Education) | .176 (*.156*) | .139 (*.128*) | .165 (*.147*) |
| Δ* | .287 (*.168*) | .279 (*.258*) | .285 (*.195*) |

participation in an education program (valid sampler) or about participation conditional on winning a prize (invalid sampler). Valid samplers were given the opportunity to sample 50 departments that either participated in the program or not and were then informed about whether they won the prize or not. Invalid samplers, in contrast, learned about the reverse contingency. That is, they could select 50 departments that either won the prize or not and received feedback about whether these departments had participated in the program.

*Individual sampling phase*

As in Experiment 2, participants could actively choose on each trial which kind of department (valid samplers) or which kind of outcome (invalid samplers) they would like to learn about. Participants playing the valid-sampler's role received the prompt "What department do you want to consider next?" in top of the screen. They could choose either "Educational program CONDUCTED" or "Educational program SUSPENDED." They then received feedback on the bottom of the screen about whether a (randomly selected) department from the chosen category had "WON the prize" or "MISSED the prize." Trials for invalid samplers started with the prompt "Which result do you want to consider next" and chose either "WON the prize" or "MISSED the prize." The feedback presented in the upper part of the screen said either "Educational program CONDUCTED" or "Educational program SUSPENDED."

At the end of the 50-trials sampling phase, all individual participants (i.e., both valid and invalid samplers) were asked to evaluate the effectiveness of the educational program in three steps. They were first asked to "imagine a randomly drawn department that participated in the program" and to indicate whether it was more likely that this department won or did not win the prize. Logically, the low base-rate (.26 or.05) of winning in all conditions calls for a high rate of negative responses. Afterwards, they made a percentage judgment of the likelihood of winning the prize when a department had participated in the educational program. Finally, they indicated the confidence of their judgment on a rating scale from 1 to 7.

*Group discussion phase*

After dyads had discussed the impact of the educational program on the likelihood of winning the prize, they jointly estimated *p* (Win Prize|Education) and *p* (Win Prize|No Education), the difference of which provided a joint estimate of Δ. Extra payment of €2 was announced for a correct choice (of Education or No Education given a positive vs. negative Δ, respectively). At the end, they were thanked, debriefed, and paid.

## 5.2 | Results and discussion

### 5.2.1 | Individual judgments

The interactive effect of (valid vs. invalid) sampling and (positive vs. negative) contingency (between education and winning the prize) on continuous individual estimates of *p* (Win Prize|Education) was significant, $F(1, 39) = 9.615$, $d = 0.993$, $p = .004$. As in preceding experiments, responses were sensitive to the sampled input; there was no sign of careless or inattentive responding. Valid samplers' individual estimates of *p* (Win Prize|Education) in the positive-Δ condition ($M = 0.309$, $SD = 0.152$) and in the negative-Δ condition ($M = 0.126$, $SD = 0.150$) resembled the objectively manipulated values of.26 and.05, respectively. The significant difference between both means, $t(39) = 3.517$, $d = 1.21$, $p_{\text{one-sided}} < .001$, testifies to participants' sensitivity and carefulness.

Invalid samplers' estimates were also sensitive to the conditional they had experienced. Although the reverse conditional probability *p* (Education|Win Prize) to which they were exposed was irrelevant to estimating *p* (Win Prize|Education), invalid samplers' estimates in the positive-Δ condition ($M = 0.686$, $SD = 0.141$) and in the negative-Δ condition ($M = 0.282$, $SD = 0.226$) were strongly biased toward the manipulated reverse conditionals of.82 and.18, respectively (see Table 2). The contrast between Δ conditions was highly significant, $t(39) = 6.949$, $d = 2.197$, $p_{\text{one-sided}} < .001$. Thus, both valid and invalid samplers based their *p* (Win Prize|Education) estimates on whatever conditional they had experienced, regardless of its logical relevance—a clear sign of MM at individual level. Because no judgments of the complementary probability *p* (Win Prize|No Education) were obtained, no individual contingencies were available, and WIO indices could not be calculated.

The subjective confidence of *p* (Win Prize|Education) estimates was significantly higher than the midpoint of the scale (4) for both invalid samplers, $M = 4.93$, $SD = 1.25$, $t(40) = 4.737$, $d = 0.744$, $p < .001$, and valid samplers, $M = 4.90$, $SD = 1.48$, $t(40) = 3.904$, $d = 0.608$, $p < .001$, despite the fact that invalid (but not valid) samplers had themselves determined by sampling how often they observed a department that had won the prize.

Note that the binary question of whether a department randomly selected for education is likely to win or not to win a price measured judges' sensitivity to the low base-rate (.26 or.05) of winning the prize regardless of education. Consistent with MM, invalid samplers provided predominantly positive responses (93.1%) in the positive-Δ condition, reflecting the high reverse conditional of *p* (Education|Win Prize) = .82, despite the fact that the high rate of A&winning cases they had observed reflected their own sampling bias. Both valid samplers (27.6% and 0%) and invalid samplers in the negative-Δ condition (8.3%), who experienced low conditionals (Table 2), were highly reluctant to provide positive responses. This interactive effect of (valid vs. invalid) sampling and (positive vs. negative) contingency (between education and prize winning) on binary estimates of *p* (Win Prize|Education) was clearly significant, $F(1, 39) = 14.494$, $d = 1.219$, $p < .001$.

### 5.2.2 | Dyadic judgments

Prior to the analysis of the dyadic contingency judgments, the estimated proportional differences $\Delta^* = p^*(\text{Win Prize}|\text{Education}) - p^*(\text{Win Prize}|\text{No Education})$ were inverted in the negative-Δ condition

such that positive Δ* scores always reflected correct judgments. Thus, the correctness criterion for every dyad is Δ = .26 − .05 = .21. The average Δ* across all dyads, $M_{\Delta^*}$ = 0.285, *SD* = 0.195, comes close to this objective criterion. The *difference* between the positive-Δ ($M_{\Delta^*}$ = 0.287, *SD* = 0.168) and the negative-Δ condition ($M_{\Delta^*}$ = 0.279, *SD* = 0.258) was negligible, $t(39) = 0.122$, $d = 0.038$, $p = .904$.

Still, MM at group level was again evident in a regression of the dyadic Δ* estimates on the *p**(Win Prize|Education) estimates provided by valid and invalid samplers. Dyads gave more weight to invalid samplers', $r = .719$, $\beta = .640$, $t(39) = 5.243$, $p < .001$, than to valid samplers' *p**(Win Prize|Education) estimates, $r = .467$, $\beta = .174$, $t(39) = 1.428$, $p = .161$. This ironic dominance of invalid opinions reflects the stronger variance of the conditionals experienced by invalid samplers. Thus, regardless of the irrelevance of invalid samplers' input and of all epistemic vigilance arising in dyadic discussions, MM governed group judgments.

## 6 | GENERAL DISCUSSION

The focus of the present research was on conflicting opinions in dyadic communication as a potential remedy to MM. MM constitutes a major source of bias and irrationality in individual judgment and decision-making (Fiedler, 2000, 2012; Fiedler, Hütter, Schott, & Kutzner, 2019, Fiedler, Hofferbert, & Wöllert, 2018, Fiedler, Schott, et al., 2019; Unkelbach et al., 2007; Juslin et al., 2007; Mahmoodi et al., 2015). Individuals' inferences often follow the given samples of information in an uncritical and naïve way. Rather than trying to separate the wheat from the chaff and to discriminate between valid and invalid observations and inferences, they often behave like "naïve samplers." They are quite sensitive to the stimulus data but fall prey to any biases that happen to be inherent in the given data. For example, in a stock-market task, judgments of the prior success of different stocks were sensitive to the frequency with which each item's success was presented on the screen but failed to distinguish between original outcomes and mere repetitions of already presented outcomes (Unkelbach et al., 2007). Or, judgments of health risks were sensitive to the estimates offered by different advisors but did not distinguish between valid and invalid advisors, whose estimates were noticeably flawed. This failure to discard invalid advice persisted even after explicit debriefing and even when participants themselves judged misleading advice to be invalid (Fiedler, Hütter, et al., 2019).

In spite of this growing evidence for MM in a variety of problem contexts and paradigms (Fiedler, 2012; Fiedler, Schott, et al., 2019; Unkelbach et al., 2007), virtually, all prior research was confined to individual performance. Hardly any investigation addressed MM in collective task settings (for a notable exception, see Bonner & Cadman, 2014). But the notion of epistemic vigilance suggests that both individual and group cognition can be improved in collective settings (Sperber et al., 2010). Adversarial discussion and conflicting opinions may alert people to the possibility of invalidity, misunderstandings, and deception and may help them to overcome their lethargic insensitivity to misleading and potentially deceptive information.

Evolutionary approaches to human interaction and communication highlight the adaptive value of epistemic vigilance (Sperber et al., 2010), conceived as a catalyst of deeper and more critical reasoning. It seems possible to enhance memory and logical inferences when cheater detection (Cosmides & Tooby, 2013) or survival motives (Nairne, Pandeirada, & Thompson, 2008) are raised. Given that participants in MM research rarely complain after debriefing that sources of bias and flawed reasoning could not be understood, it seems plausible that conflict-prone discussions might alert them for biases in the stimulus input. Collective settings suggest a well-reasoned remedy to MM. Or, conversely, persistent biases in spite of the epistemic vigilance raised in social settings would present particularly strong evidence for MM.

Conversely, persistent biases in spite of the epistemic vigilance raised in social settings would present particularly strong evidence for MM. For a straightforward test, we engaged randomly paired dyads in a joint lottery decision task, which called for comparative estimates of the winning probabilities *p* (win|A) and *p* (win|B) and a choice between two Lotteries A and B. Whereas one member of the dyad, the valid sampler, received outcome feedback (win vs. not win) conditional on the lottery sampled, the invalid sampler received feedback about the lottery (A vs. B) conditional on the outcome chosen. Because samples in the latter condition were biased toward the lottery base rates *p* (A) and *p*(B)—that is, *p*(A|win) must be higher/lower than *p*(B|win) when *p*(A) is much higher/lower than *p*(B)—the invalid sampler's stimulus input was severely biased. Only valid samples provided unbiased estimates of the lotteries' success chances. The crucial question was whether a discussion between valid and invalid samplers would induce epistemic vigilance and enable dyads to jointly overcome MM, that is, to follow valid samples and to discard invalid samples.

How could this be accomplished? Applying the notion of demonstrability to dyadic choices suggests that the joint response could be expected to be correct either when both group members hold the correct solution or when only one member (i.e., the valid sampler) infers the correct solution and understands, at the metacognitive level, that he/she must persuasively communicate this solution to the other member (i.e., the invalid sampler). Although invalid sampler's individual MM prevents the former case, the latter possibility suggests that MM may be overcome through effective communication.

Otherwise, MM may persist in dyadic group settings, and epistemic vigilance may not be sufficient to prevent dyads from assigning substantial weights to, or even average, estimates derived from both samples. Invalid samplers' individual estimates provided a measure of MM at the individual level, whereas dyadic judgments and choices reflected MM at the collective level. A postexperimental questionnaire assessed whether group discussions were reflective of strategies consistent with the notion of epistemic vigilance.

The empirical results obtained in all three reported experiments converge in providing strong and consistent evidence for MM at both levels. Although the questionnaire responses showed that the dyads' discussions did revolve around such epistemic-vigilance topics as discrepancy, validity, logical status, and information value of sample, the joint comparative evaluations of *p* (win|A) and *p* (win|B) were clearly

biased toward invalid samplers' estimates. Even though many dyads did prioritize valid over invalid samples, their final collective judgments represented an unwarranted compromise of both sources, reflecting a substantial weight given to invalid samplers' misleading evidence. Ironically, because there was more systematic variation in invalid than in valid samples, due to the *p*(A|win) manipulation, the invalid samples became the major determinant of the final group performance.

Thus, although the discussions were not too superficial to prevent the dyads from noting the discrepancy of valid and invalid samples or to fully ignore the validity of the arguments, any impact of epistemic vigilance was apparently overshadowed by the uncritical tendency to give roughly equal weight to every opinion. Apparently, then, group dynamics and communication skills did not help dyads to overcome individual members' metacognitive confines.

A similar tendency to give too much weight to invalid opinions that were recently observed in an advice-taking situations (Fiedler, Hütter, et al., 2019; Mahmoodi et al., 2015) suggests that maybe the motive to get along with others (Snyder, 1992; Wittenbaum, Hollingshead, & Botero, 2004) creates an equality bias that may counteract and overshadow the impact of epistemic vigilance. In any case, our findings testify to strong and persistent MM at the dyadic group level. Group discussions not only fail to correct for individual-level biases. They actually reinforce the bias when invalid inferences vary more strongly than valid inferences. Before dyadic deliberation came into play, MM at the individual level prevented invalid samplers from noting the misleading nature of their reverse information samples.

Everyday examples show that reversed conditional probabilities can deviate dramatically. Of course, the conditional probability *p* (fever|malaria) is much higher than the reverse conditional *p* (malaria|fever). Invalid like valid samplers (in the subsequent discussion) should thus have had a fair chance—and even an intellectual obligation—to note that *p*(A|win) must not be mistaken for *p* (win|A). It would be cynical to argue that smart adult psychology students lack the capacity to discriminate between valid and invalid samples—provided they really make an attempt to overcome MM and engage in critical assessment.

However, beyond admitting that task difficulty did not exceed normal intelligence or numeracy (Reyna & Brainerd, 2008), we also established a situation in which "invalid" samples were not absolutely "invalid," as they did contain all information required to make unbiased estimates of the task-relevant quantities, *p* (win|A) and *p* (win|B). Whereas invalid samplers in Experiment 1 only received evidence on lotteries given winning outcomes but no information about lotteries associated with not winning—motivated by the intention to make invalid sampling useless or "surreal"—participants in Experiments 2 and 3 were exposed to the entire 2 × 2 distribution of lotteries and outcomes. As a consequence of this change in procedure, all participants could assess joint frequencies of all 2 × 2 combinations, which allowed both valid and invalid samplers to produce unbiased estimates of *p* (win|A) and *p* (win|B): They only had to divide the joint frequency (A&win) by the summed joint frequencies of (A&win) + (A&not win) and subtract the frequency ratio of (B&win) and (B&win) + (B&not win). However, rather than following such a fail-proof joint-frequency strategy, they uncritically relied on whatever conditional samples they were exposed to. Valid samplers relied on winning feedback given sampled lotteries, whereas invalid samples relied on lottery feedback given sampled outcomes.

Experiment 3 extended and corroborated these findings in the context of a new task, replacing lotteries with prize winning after an educational intervention. In this modified task, the impact of MM (i.e., the manipulation of how often winning a prize was preceded by an educational program) had to compete with one-sided prior expectations (that the chances to win a prize will presumably increase rather than decrease after an educational intervention). However, these a priori constraints did not prevent naïve samplers from being misled by the superficial association of prize winning and education in the invalid sampling condition.

It would be interesting to find out whether and under what conditions groups may be more sensitive to discriminating valid and invalid information sources. The seminal work of Laughlin and colleagues (Laughlin et al., 2008; Laughlin & Ellis, 1986) suggests that on intellective problems with a clearly identifiable, demonstrably correct solution, groups will outperform individuals only under distinct conditions. Granting the existence of a unique solution, a social combination mechanism (Laughlin & Adamopoulos, 1980; Laughlin & Ellis, 1986) implies that group reasoning will converge on the true answer only if a sizable subset of group members understands and effectively persuades the remaining group members of the correct solution. So the long-established work on demonstrability already anticipated an answer to the present finding that epistemic vigilance may not be an effective remedy to MM.

Although this intriguing hypothesis deserves to be tested in future research, it seems unclear whether the statistical sampling tasks that give rise to MM deficits meet the premises of a social combination approach. Although our lottery choice task renders "valid sampling" incontestably adequate, and "invalid sampling" demonstrably misleading, the correct solution here is not a fixed numerical or verbal response option as in the aforementioned literature. Proper sampling depends on an abstract principle (such as conditional probabilities) that may be hard to be communicated and negotiated in groups. It may thus turn out that demonstrability may not be enough; groups may only overcome MM when a convergent analytical problem renders the singular solution easy to communicate.

Another potentially effective intervention that may help to decrease MM is to render group discussion contingent on knowledge transfer, whereby group members are instructed to reflect explicitly on their relevant knowledge before they contribute to group discussion and problem solving (Bonner & Baumann, 2012).

At the end, the empirical progress in our attempt to find a social remedy to MM is far from a "happy end" (see also Fiedler, Hütter, et al., 2019; Fiedler, Schott et al., 2019; Fiedler, McCaughey, Prager, Eichberger, & Schnell, 2019). But although we remain optimistic and confident to find an effective debiasing procedure under auspicious conditions (cf. Mata, Fiedler, Ferreira, & Almeida, 2013)—and we strongly applaud to any demonstration that this is possible—the available evidence seems to underpin the robustness rather than the

curability of MM (Fiedler, 2008, 2012; Unkelbach et al., 2007). MM may turn out to be as serious a source of irrationality as capacity constraints, misunderstandings of instructions, motivated biases, and other causes that have been the focus of extended experimental research.

## ACKNOWLEDGEMENT

The research underlying the present paper was supported by a Koselleck grant and another grant (Fi 294/23-1) and (Fi 294/26-1) awarded by the Deutsche Forschungsgemeinschaft to the first author. Original materials can be retrieved from https://www.psychologie.uni-heidelberg.de/ae/crisp/studies/ultidyad.html.

## ORCID

*Klaus Fiedler* https://orcid.org/0000-0002-3475-0868

## REFERENCES

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Bonner, B. L., & Baumann, M. R. (2012). Leveraging member expertise to improve knowledge transfer and demonstrability in groups. *Journal of Personality and Social Psychology*, *102*(2), 337–350. https://doi.org/10.1037/a0025566

Bonner, B. L., & Cadman, B. D. (2014). Group judgment and advice-taking: The social context underlying CEO compensation decisions. *Group Dynamics: Theory, Research, and Practice*, *18*(4), 302–317. https://doi.org/10.1037/gdn0000011

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, *64*, 201–229. https://doi.org/10.1146/annurev.psych.121208.131628

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*(3), 399–418. https://doi.org/10.1002/ejsp.168

Fiedler, K. (2000). Beware of samples: A cognitive-ecological sampling approach to judgment bias. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037/0033-295X.107.4.659

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *34*(1), 186–203. https://doi.org/10.1037/0278-7393.34.1.186

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57) (pp. 1–55). San Diego, CA, US: Elsevier Academic Press.

Fiedler, K., Hofferbert, J., & Wöllert, F. (2018). Metacognitive myopia in hidden-profile tasks: The failure to control for repetition biases. *Frontiers in Psychology*, *9*, 903. https://doi.org/10.3389/fpsyg.2018.00903

Fiedler, K., Hütter, M., Schott, M., & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*, *32*(3), 317–333. https://doi.org/10.1002/bdm.2109

Fiedler, K., McCaughey, L., Prager, J., Eichberger, J., & Schnell, K. (2019). Speed-accuracy tradeoffs in sample-based choices. Manuscript submitted for publication.

Fiedler, K., Schott, M., Kareev, Y., Avrahami, J., Ackerman, R., Goldsmith, M., … Pantazi, M. (2019). Metacognitive myopia in change detection: A collective approach to overcome a persistent anomaly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000751

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, *11*(4), 401–415.

Hütter, M., & Fiedler, K. (2019). Advice taking under uncertainty: The impact of genuine advice versus arbitrary anchors on judgment. *Journal of Experimental Social Psychology*.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678–703. https://doi.org/10.1037/0033-295X.114.3.678

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Larson, J. R., Foster-Fishman, P. G., & Keys, C. B. (1994). Discussion of shared and unshared information in decision-making groups. *Journal of Personality and Social Psychology*, *67*(3), 446–461. https://doi.org/10.1037/0022-3514.67.3.446

Laughlin, P. R., & Adamopoulos, J. (1980). Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology*, *38*(6), 941–947. https://doi.org/10.1037/0022-3514.38.6.941

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations*, *11*(3), 319–330. https://doi.org/10.1177/1368430208090645

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, *90*(4), 644–651. https://doi.org/10.1037/0022-3514.90.4.644

Laughlin, P R. (2011). Social choice theory, social decision scheme theory, and group decision-making. *Group Processes & Intergroup Relations*, *14*(1), 63–79.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., … Roepstorff, A. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840. https://doi.org/10.1073/pnas.1421692112

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*(3), 486–491. https://doi.org/10.1016/j.jesp.2013.01.010

Nairne, J. S., Pandeirada, J. S., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*(2), 176–180. https://doi.org/10.1111/j.1467-9280.2008.02064.x

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology*, *23*(1), 305–343.

Snyder, M. (1992). Motivational foundations of behavioral confirmation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25) (pp. 67–114). San Diego, CA US: Academic Press. https://doi.org/10.1016/S0065-2601(08)60282-8

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stewart, N., Chater, N., & Brown, G. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Unkelbach, C., Fiedler, K., & Freytag, P. (2007). Information repetition in evaluative judgments: Easy to monitor, hard to control. *Organizational Behavior and Human Decision Processes*, *103*(1), 37–52. https://doi.org/10.1016/j.obhdp.2006.12.002

Wittenbaum, G. M., Hollingshead, A. B., & Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. *Communication Monographs*, *71*(3), 286–310. https://doi.org/10.1080/0363452042000299894

Wittenbaum, G. M., & Park, E. S. (2001). The collective preference for shared information. *Current Directions in Psychological Science*, *10*(2), 70–73. https://doi.org/10.1111/1467-8721.00118

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory & Cognition*, *35*(2), 558–563. https://doi.org/10.1037/a0014589

**How to cite this article:** Fiedler K, Krüger T, Koch A, Kutzner F. Dyadic judgments based on conflicting samples: The failure to ignore invalid input. *J Behav Dec Making*. 2020; 1–13. https://doi.org/10.1002/bdm.2173

## APPENDIX A

Postexperimental questionnaire supposed to assess the degree to which the discussions preceding the final dyadic judgments reflected metacognitive awareness of sample discrepancies. Items 2, 3, 4, 5, 6, and 8 afford a measure of epistemic vigilance.

1. How long approximately did your discussion last?

| | | | |
|---|---|---|---|
| <1 min ① | 1–3 min ① | 3–5 min ② | >5 min ④ |

2. How much agreement did you finally reach about what was the better lottery?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| No agreement at all | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | Full agreement |

3. The samples observed by both partners originally suggested lottery evaluations that were

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Very similar | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | Incompatible |

4. How strong was the discrepancy of evaluations that had to be reconciled?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Very strong | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | Not existent |

5. Did you give equal weight to both individual evaluations or did you follow one partner's evaluations who you recognized to be correct?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Equally weighted | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | One partner |

6. What determined the joint evaluations more, the relative eagerness of the two discussants or the information value of their samples?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample value | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | Eagerness |

7. Whose sample was logically more suitable for the evaluation of lotteries?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| My sample | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | The other sample |

8. Was the logical status of samples a genuine discussion point between the two of you?

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Definitely no | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | Definitely yes |