**Metacognitive Myopia in Change Detection:
A Collective Approach to Overcome a Persistent Anomaly**

Klaus Fiedler, Malte Schott (University of Heidelberg), Yaakov Kareev, Judith Avrahami (Hebrew University), Rakefet Ackerman (Technion – Israel Institute of Technology; University of Haifa), Morre Goldsmith (University of Haifa), André Mata, Mário Ferreira (University of Lisbon), Ben Newell (University of New South Wales), and Myrto Pantazi (University of Cambridge)

Running head: Metacognitive myopia in change judgments

Word count: 13263 words

Author Note: The research underlying this article was supported by a grant provided by the Deutsche Forschungsgemeinschaft to the first author (FI 294/26-1). Helpful comments by Ola Svenson are gratefully acknowledged. Email correspondence can be addressed to kf@psychologie.uni-heidelberg.de

Abstract

Going beyond the origins of cognitive biases, which have been the focus of continued research, the notion of metacognitive myopia refers to the failure to monitor, control, and correct for biased inferences at the metacognitive level. Judgments often follow the given information uncritically, even when it is easy to find out or explicitly explained that information samples are misleading or invalid. The present research is concerned with metacognitive myopia in judgments of change. Participants had to decide whether pairs of binomial samples were drawn from populations with decreasing, equal, or increasing proportions p of a critical feature. Judgments of p changes were strongly affected by changes in absolute sample size n , such that only increases (decreases) in p that came along with increasing (decreasing) n were readily detected. Across four experiments these anomalies persisted even though the distinction of p and n was strongly emphasized through outcome feedback and full debriefing (Experiment 1-4), simultaneous presentation (Experiments 2-4) and recoding of experienced samples into descriptive percentages (Experiment 3-4). In Experiment 4, a joint attempt was made by ten scientists working in seven different institutions to develop an effective debiasing training, suggesting how multi-lab-collaboration might improve the quality of science in the early stage of operational research designing. Despite significant improvements in change judgments, debiasing treatments did not eliminate the anomalies. Possible ways of dealing with the metacognitive deficit are discussed.

Introduction

The present research is concerned with metacognitive myopia (MM), conceived as a serious restriction in the metacognitive functions required to monitor the quality, and to control and correct for biased judgments and decisions. Note, for the sake of terminological precision, that our use of the term “metacognitive” refers in a broad sense to all mental functions involved in monitoring and control (Johnson-Laird, Miller, Galanter, Pribram, Nelson, & Narens, 1992). We adopt Nelson’s (1996) view that “empirical investigations of metacognitive monitoring and its effect on individuals’ metacognitive control can give us important clues about both what people know ... and why they choose to behave as they do” [p. 114]. The term “myopia” suggests that judgments tend to be short-sighted, doing justice to the proximal information given in a stimulus sample but failing to make appropriate inferences about the distal world from which the sample is drawn. Thus, metacognitive myopia refers to a deficit in monitoring and control.

A growing body of evidence (Fiedler, 2000; 2012) suggests that judgments are often remarkably sensitive to even complex stimulus samples (e.g., different performance aspects of many students in a simulated school class; Fiedler, Woellert, Tauber & Hess, 2013) but uncritical and naïve regarding the validity of the given sample. Taking sampled information for granted, myopic judges hardly notice when systematic biases render samples invalid and they often forego the possibility to control and correct for upcoming biases, even when they are sensitized to the problem or debriefed explicitly (Fiedler, Hofferbert & Wöllert, 2018; Fiedler, Hütter, Schott & Kutzner, 2019; Unkelbach, Fiedler & Freytag, 2007). Indeed, the failure to correct for preliminary biases constitutes a major impediment to rational behavior (Fiedler, 2008, 2012).

Demonstrations of MM vary on a continuum from subtle and hard-to-avoid to blatant and easily understood biases. For example, irrational consequences of the base-rate fallacy may, on the one hand, reflect a genuine difficulty to understand the logical necessity to take base-rates into account. On the other hand, MM continues to distort judgments when sampling biases are obvious and even when

judges are explicitly debriefed and warned not to use a biased sample (as in the experiments reported below). For example, preferences in personnel selection were systematically biased by the mere repetition of some applicants' assets and other applicants' deficits, even though repetitions were fully redundant and participants were explicitly warned to withstand the distorting influence of selective repetition (Fiedler et al., 2018). When the task in a stock-market game was to assess how often various shares were among the daily winners, judgments were biased toward those shares whose winning outcomes had been presented twice, even after a warning not to attend to (potentially misleading) repetitions (Unkelbach et al., 2007). Estimates of health risks were influenced by an advice giver whose estimate was invalid, even when participants were fully debriefed about the invalidity of the advice (Fiedler et al., 2018).

Metacognitive Myopia in Detection of Change

The focus of the present research is on MM in judgments of change – a task setting that involves sampling biases of intermediate transparency. On every trial of a sequential paradigm, participants experience a random sample of sequentially presented binary symbols (e.g., smiling and frowning faces). They have to decide whether the proportion p of a focal symbol (e.g., smiling faces) in the universe from which the current sample is drawn is the same as the universe from which the preceding sample was drawn, or whether the current sample stems from a new universe with a higher or lower p . Note that this task is not confined to estimating and comparing manifest sample proportions. As in many areas of real life (stock market; sales; student performance), the aim of change detection is to make inferences from observed samples to latent populations.

Although the task focus is clearly on the relative proportion p , change detection is seriously biased when successive samples also vary in absolute size n (Fiedler, Kareev, Avrahami, Beier, Kutzner & Hütter, 2016). An increase (decrease) in p is only detected readily when n also increases (decreases). Conversely, changes are hard to detect when increases (decreases) in p come along with decreases (increases) in n . When p remains invariant, increasing (decreasing) n induces erroneous

increase (decrease) judgments. Thus, an increase from 4/8 (smileys/faces) to 12/16 is readily noticed, just as a decrease from 6/16 to 2/8, but the same proportional changes are hard to detect when 8/16 increases to 6/8 or when 3/8 decreases to 4/16. Given equal p , 10/16 after 5/8 appears like an increase but 7/8 after 14/16 appears like a decrease. Here MM consists in the inability to set changes in relative proportions apart from changes in absolute sample size.

Everybody understands that quantity does not imply quality. Large groups need not be better than small groups; those who speak most need not have the best arguments; or a high citation index (visibility) must not be confused with quality of science (validity). Indeed, an ecological analysis reveals an opposite, negative correlation between value and frequency of occurrence (Pleskac & Hertwig, 2014); market prices and aesthetic values are higher for scarce than for frequent objects. And yet, the confusion of extensional (n) and intensional (p) information is conspicuously robust and common, not only in change judgments (Fiedler et al., 2016), but also in other well-established paradigms such as the ratio bias (Kirkpatrick & Epstein, 1992), the denominator neglect (Reyna & Brainerd, 2008), a popularity bias in consumer choice (Powell, Yu, DeWolf & Holyoak, 2017), in numerical quantitative priming (Oppenheimer, LeBouf & Brewer, 2008), or sample-size biases in risk estimation (Price, 2001; Price, Smith & Lench, 2006) and additive calculation tasks (Smith & Price, 2010). Even when participants in such experiments are motivated to be accurate and to increase their performance-contingent payment, they hardly manage to stop confusing n and p .

A viable account. Although the persistent difficulty to assess proportions (p) independently of sample size (n) appears to reflect a serious category mistake, one might explain the anomaly in terms of natural constraints imposed on the cognitive assessment process. Imagine an experiment that calls for the inductive assessment of proportions p of smiling faces in sequentially presented binary samples, the size of which is the summed frequency (f) of smiling and frowning faces. Thus, $n = f_{\text{smile}} + f_{\text{frown}}$ and $p = f_{\text{smile}} / (f_{\text{smile}} + f_{\text{frown}}) = f_{\text{smile}} / n$. Let us assume that the smiling rate in the universe from which the sample is drawn is quite high, say 75%; three quarters of all sampled observations can be expected to

be smiling faces. We understand that in order to estimate $p = f_{\text{smile}}/n$, the weight given to each elementary observation is $1/n$. In a very small sample of size 2, the $1/2$ weight of both observations is rather high; in a sample of 10, the $1/10$ weight is smaller; in a large sample of 100, the $1/100$ weight of each observation is minimal.

However, in an experiment with multiple trials of varying n – just as in many real-life tasks – n is not known beforehand. We therefore do not know the increment or decrement needed to update the p estimate after smiling or frowning faces, respectively, for the updating factor ($1/n$) is uncertain. As a consequence, the resulting p estimates will, to an unknown degree, reflect a summation rule rather than an averaging rule, or some compromise of summation ($p \cdot n$) and averaging (p), which is at least partly contaminated with n .

A sign of moderation. Consistent with this notion, Fiedler et al. (2016) found anomalies in change detection only when proportions had to be extracted in an inductive-statistical inference task from sequentially experienced samples of elementary observations. Change judgments were unbiased when the same stimulus samples were presented descriptively, that is, when sample proportions were described in percentage format (e.g. a change from 75% in a sample of 8 to 87% in a sample of 16).¹ In a similar vein, it is easy to see whether the proportion of colored area in a pair of pie charts increases or decreases, regardless of whether the total pie area (or diameter) increases or decreases. Pie chart areas, like stated probabilities, can be perceived directly; they do not have to be inferred inductively from elementary events in a continuous updating process.²

This brief discussion of moderating task conditions is only meant to rule out suspicion that setting p apart from n may be inherently too difficult or participants may not understand the task of

¹ Note that this finding is opposite to the common notion that normalized probability formats lead to less accurate judgments than natural frequency formats (Hoffrage & Gigerenzer, 1998).

² We thank an unknown reviewer for suggesting this example.

inferring latent changes from given samples. It does not restrict the domain of the phenomenon being studied to an artificial task format. Rather, the kind of experienced samples that were shown to produce the anomaly are at the heart of many change-judgment tasks in real life, such as assessments of changing student achievement, sales rates, voting outcomes, or sports results.

Moreover, a plethora of related findings demonstrate that the anomaly generalizes to other task formats, as evident from ratio bias, denominator neglect, sample-size bias in risk estimation and many other tasks (Denes-Raj & Epstein, 1994; Fiedler et al., 2016; Reyna & Brainerd, 2008). Thus, even though the aforementioned account in terms of the unknown factor $1/n$ in sequential updating may suggest a possible learning origin, the anomaly seems to generalize to many other task formats, maybe reflecting some generalized learned-helplessness effect (Mikulincer, 2013) due to n uncertainty experienced under many natural conditions.³ In any case, the available evidence suggests that even explicitly known n can intrude into judgments that ought to abstract from n (e.g., additive calculation of n constituents; Smith & Price, 2010; frequency of Amazon ratings overriding average rating of consumer products, Powell et al., 2017).

Getting the study purpose right. Yet, importantly, the purpose of the present research is definitely not to unravel the specific conditions that give rise to the anomaly or to explain an underlying cognitive mechanism that produces the anomaly. The present research is rather concerned with the metacognitive monitoring and control stage that should support correction of any bias – whatever its origin – in the light of feedback and debriefing. Granting that cognitive biases are ubiquitous, rationality may still be reestablished at the metacognitive level. Relying on an established paradigm that was shown to produce a distinct anomaly in change detection, we investigate the metacognitive ability to overcome this anomaly in the light of increasingly explicit feedback about the

³ Another explanation for the generality of the bias might rely on the assumption that n is more or less constant in reality, because working memory restricts effective samples to the same virtual size.

misleading impact of n on p assessment. This research goal is independent of a mechanistic account of why and under what conditions the anomaly arises in the first place.

Such a perspective diverges from traditional approaches that treat fallacies and biases as first impressions. Like Hertwig and Grüne-Yanoff (2017), we believe “there is more to human decision making and problem solving than this first response” [p. 975]. Granting that cognitive biases are ubiquitous and impossible to avoid in the first place, rational behavior relies heavily on metacognition (i.e., monitoring and correction for bias; Ackerman & Thompson, 2017; De Neys & Bonnefon, 2013; Mata, Ferreira & Sherman, 2013; Stanovich, 2011; Thompson, Turner & Pennycook, 2011), independent of why upcoming biases were not prevented from the beginning.

From such a metacognitive perspective on rationality as an obligation to engage in monitoring and control of cognitive inferences, an intriguing question is whether MM can be overcome through distinct feedback and debiasing training. Tackling this question within the depicted change-judgment paradigm, we conducted a series of experiments on how to overcome the intrusion of absolute n -changes into the assessment of proportional p -changes. We believe that the findings obtained in this paradigm have broad implications for metacognitive training in rational decision and action, highlighting MM as an impediment to rational behavior.

Another, no less engaging goal was to realize this research in a multi-lab collaboration involving expert scientists who all share an interest in metacognitive research but who vary in the optimistic belief how easy it is to overcome MM using appropriate training procedures. The purpose here is to highlight that the advantage of multi-lab collaborations is not confined to increasing the power of statistical analyses. Transcending the confines of single labs can also enrich and improve the quality of design and operationalization in early stages of research.

Preview of Experiments

The remainder of this article is devoted to experiments involving a variety of debiasing interventions. The aim of the first experiment is to outline the basic set-up and to replicate the pattern

of anomalies obtained by Fiedler et al. (2016), as a premise to all subsequent attempts to locate and to undo the deficit. As in previous studies, Experiment 1 includes p -changes (i.e., decreasing, equal, or increasing sample proportions) and n -changes (decreasing vs. increasing sample sizes) as orthogonal repeated-measures factors. The task calls for judgments of change between successive samples; one sample is currently visible while the preceding sample has to be kept in memory. In addition to replicating this task in a control condition, two other conditions involve increasing feedback information. One condition receives, after the first trial block, an explicit debriefing and a warning not to be misled by n -changes when judging p -changes; the other condition includes the same warning plus outcome feedback of the correctness of change judgments on each trial. Both levels of feedback should alert participants for the confusion of p and n and should thereby help them to avoid the mistake.

A similar design is used in Experiment 2, with two modifications. First, only two groups are included, a no-feedback control group and a full-feedback condition. Second, the two stimulus samples are now (as in all later experiments) presented simultaneously on the same screen, eliminating memory load and facilitating comparative encoding of sample proportions. This simultaneous format also greatly facilitates linking correctness feedback to p and n in an exact copy of both samples. It is hard to see how such vivid feedback should not help to set p apart from n .

Experiment 3 includes the same two within-participants factors (p -changes $\times$ n -changes) along with two between-participants manipulations. Orthogonal to the same no-feedback versus full-feedback manipulation as in Experiment 2, one subgroup in both conditions is asked to provide percentage estimates of both sample proportions before the final change judgment, whereas no such estimates are provided by the other subgroups. Because normalized percentage format had erased the anomalies in prior research (Fiedler et al., 2016), we want to check on the possibility that self-generated percentage estimates afford an effective prompt for debiasing.

To anticipate the persistent result, the metacognitive confusion of p and n remained strong and largely unaffected by all these debiasing attempts. What appeared to be strong and effective

interventions did not really help participants to overcome the intrusion of n changes in judgments of p changes. We finally resorted to a deliberate attempt to devise the most effective debiasing training that a number of expert scientists in the field could collectively think of. We invited decision researchers in diverse labs to help us create an effective debiasing treatment. This multi-lab-collaboration will be described in more detail in the introduction to the final Experiment 4. For the moment, though, we first consider the results of the three experiments previewed so far.

Experiment 1

As already mentioned, we started with an extended replication of the basic phenomenon under different feedback conditions, asking for change judgments in successive samples.

Methods

Participants and Design. Seventy-five male and female Heidelberg students participated either for payment (6 € for a 45 minutes session) or to meet a study requirement. They were randomly assigned to one of three (equal-sized) experimental groups: no feedback, debriefing, or debriefing + trial-by-trial feedback. Within participants, three levels of p -changes (decreasing, equal, increasing) and two levels of n -changes (decreasing, increasing) varied across trials.

Materials and procedures. The entire experiment was conducted in computer dialog. Initial on-screen instructions explained that the present research was "... concerned with the detection of changes in the environment. It is not always easy to discriminate real changes from random fluctuations. Did the weather or the climate actually change or only exhibit normal variations?" After a few examples, participants were told that the task was developed to study these distinctions. "In each trial of a longer series, you will be shown a sample of two symbols, "≡" and "o", appearing at varying frequencies. The sample is randomly drawn from a universe of 100 symbols, in which the rate of both symbols is clearly determined. The question on each trial is: Was the current sample drawn from the same universe as the preceding sample or did the universe from which the new sample stems undergo a change?"

The purpose of the following, redundant passage was to accustom the participant to the jargon of the change-detection task: “Was the last sample drawn from the same ‘world’ as the second-last sample, that is, could the varying rate of the two symbols be due to chance, reflecting merely another sample drawn from the same ‘world’? Or must the ‘world’ have changed, that is, must the last sample have been drawn from a changed universe in which the rate of both symbols is no longer the same?”

The remainder of the instruction text was largely identical to the following description of the sequential procedure. It was explained to participants that the critical question on each trial would always focus on one focal symbol (i.e., $\equiv$). They would be asked to judge whether the proportion of $\equiv$ symbols in the ‘world’ has changed or not. To provide their answer, they were told to press one of three response keys they would be offered (see Figure 1). The next trial would then follow shortly afterwards. Moreover, there would be three blocks of such trials. After two warm-up trials, they were explicitly instructed to provide as many correct responses as possible, and they were reminded of the fact that an objective correctness criterion exists as each sample is drawn either from the same or from a changed universe. They were promised a detailed correctness record at the end of the experiment.

The binary symbols comprising each sample were presented successively at a rate of one symbol per 100 ms in random order and in random locations of a 480 x 400-pixel black rectangular frame, as shown in Figure 1. Each sample contained either $n = 8$ or $n = 16$ symbols. Upon completion, the full sample remained onscreen for another 1000 ms. All symbols were then erased and participants were instructed to use three response keys to indicate that the current sample was drawn from a ‘world’ with EQUAL ($\downarrow$), LESS ($\leftarrow$), or MORE ($\rightarrow$) $\equiv$ symbols than in the previous trial. Immediately after the participant’s choice, the next trial started with the appearance of a new empty black sample frame, followed by the next sample after 1000 ms.

Small sample ($n = 8$) presented on previous trial at a rate of one symbol per 100 ms in random order and in random locations in a black rectangular frame surrounded by a light-brown screen

Large sample ($n = 16$) presented on the current trial in the same format

Current sample is erased 1000 ms after the last element is presented, and response instruction prompt is inserted on top of the

Figure 1: Schematic representation of three successive stages of the change-judgment task.

Each block consisted of 36 trials, 12 of which were no-change trials (i.e., samples drawn from the same universe), 12 involved decreases, and 12 involved increases in the proportion of the focal symbol “≡”. More precisely, in the universe, from which the samples were drawn at random (using the same algorithm as Fiedler et al., 2016), increases or decreases amounted to $\Delta p = +.20$ or $-.20$ for six trials and to $+.10$ and $-.10$ on the remaining six trials. As each new sample became the preceding sample on

the next trial, a block can be understood as a polygon of p -values with 12 zero transitions, 12 downward and 12 upward transitions. Half of the transitions at each p -change level involved a decrease in sample size from $n = 16$ to $n = 8$, whereas the other half involved an increase from $n = 8$ to $n = 16$. The sequence of polygon changes was random, subject to the constraint that p cannot exceed a lower limit of 0 and an upper limit of 1.0.

At the end of the first trial block, the feedback manipulation was introduced. Participants in the no-feedback condition simply read, “So, this was the first block of trials ... Rest a moment and then use any keystroke to start the second block of trials involving the same task.” The same text was provided in the partial-feedback condition, followed by an explicit debriefing:

“Make an attempt to improve your performance by avoiding the following mistake. It turned out that most participants are misled by the size of the samples. They recognize increases (in the proportion of the critical symbol) only when the absolute number of all symbols increases but not when the absolute number decreases. Conversely, they recognize decreases only when the absolute number of elements decreases but not when the absolute number increases. MAKE A DELIBERATE ATTEMPT TO AVOID THIS MISTAKE!”

Additionally, participants in the full-feedback condition were told, “What should be of help for you toward this end is that after each decision you will receive feedback on what would have been the correct response. The correct option will be depicted in green”. Consistent with this instruction, participants in the full-feedback condition were provided with verbal feedback (in light green letters) saying that the current sample was drawn from a world with a higher, lower or equal rate of focal $\equiv$ -symbols.

Between the second and third trial block, participants in all conditions were prepared for a change in the stimulus materials: “Let us finally see whether there is evidence for learning, that is, whether you have improved your ability to detect changes over the preceding two trial blocks and

whether this learning carries over to new materials.” They were then explained that the last block would involve the same task, but without feedback and with the hitherto used symbols (“ $\equiv$ ” and “o”) replaced by new symbols, “ $\star$ ” and “ Δ ”, respectively. The task would now be to decide whether the proportion of “ $\star$ ” has changed or not. When this last trial block was over, they were thanked and provided with a table showing their performance across all three blocks of 36 trials.

Results and Discussion

Stage 1. An analysis of the change judgments provided during the first trial block, prior to the introduction of the feedback treatment, replicated the typical anomalies in virtually all detail. The results are reported as mean change judgments on a scale from -1 (coded for “Decrease” responses) to 0 (for “Equal” responses) to $+1$ (for “Increase” responses), pooling across $.10$ and $.20$ changes. For convenience, we analyze the performance at Stage 1, 2 and 3, consecutively, with a view on possible feedback and learning effects across trial blocks.

Figure 2: Mean change judgments in Stage 1 of Experiment 1 as a function of p -change and n -change (pooling across feedback conditions). Error bars indicate standard errors of the mean.

Figure 2 shows that in the first trial block, prior to the feedback treatments, the mean change judgment scores (pooled across feedback conditions) clearly increase from left to right, reflecting

accuracy motivation and sensitivity to actually manipulated changes. A corresponding main effect for p -changes is highly significant, $F(2,148) = 72.447, p < .001, \eta^2 = 0.635$.⁴ Yet, a comparison of black and shaded bars also reflects a strong main effect for n -changes, $F(1,74) = 74.967, p < .001, \eta^2 = 0.307$. At all levels of p -change, judgments exhibit a strong and regular bias toward “increase” (“decrease”) on trials involving increasing (decreasing) samples size n . The interaction was also significant, $F(2,148) = 18.973, p < .001, \eta^2 = 0.204$, due to a stronger sample-size effect for no-change trials (see middle pair of bars in Figure 1).

This bias was so strong that average change judgments at all three p -change levels almost completely fall in the negative scale range when n decreased but in the positive scale range when n increased. A majority of 63 out of all 75 participants (84%) produced on average higher p -change judgments on increasing- n trials compared to decreasing- n trials, reflecting a high generality of the sample-size intrusion across individual participants.

Stage 2. Having replicated the basic phenomenon in the first stage, the next question that suggests itself is whether the intrusion of absolute n -changes into judgments of relative p -changes was ameliorated through debriefing and feedback. So let us consider the performance in the second trial block, in which the vivid feedback experience and the debiasing instructions should have improved the change-judgment performance immediately.

It is no surprise that the no-feedback control condition produced more or less the same results: a highly significant main effect for p -changes, $F(2,48) = 15.463, p < .001, \eta^2 = .629$, a strong main effect for n -changes, $F(1,24) = 20.886, p < .001, \eta^2 = .206$, and no interaction, $F(2,48) = 2.242, p = .117, \eta^2 = .085$. Change judgments increased similarly from decreasing to equal to increasing trials when n either decreased ($-0.523, -0.140, -.183$) or when n increased, though at a higher level ($0.073, 0.267, 0.317$).

⁴ Sensitivity to p -changes was even stronger in alternative ANOVAs that reduced the p -changes factor to a linear trend with $df = 1$, which amounts to giving a zero weight to no-change trials.

More surprisingly, though, the same basic pattern was also maintained in both feedback conditions. With partial feedback, the mean change judgments for the three p -change conditions amounted to -0.369 , -0.200 , -0.062 for decreasing n and 0.187 , 0.333 , 0.493 for increasing n , thus yielding main effects for p -change, $F(2,48) = 18.030$, $p < .001$, $\eta^2 = .671$, a strong main effect for n -changes, $F(1,24) = 26.086$, $p < .001$, $\eta^2 = .185$, and no interaction, $F(2,48) = 0.026$, $p = .974$, $\eta^2 = .001$. With full feedback, the main effects for p -change, $F(2,48) = 22.876$, $p < .001$, $\eta^2 = .443$, and for n -changes were similarly strong, $F(1,24) = 22.452$, $p < .001$, $\eta^2 = .269$, and no interaction was obtained, $F(2,48) = 2.714$, $p = .076$, $\eta^2 = .102$.

When feedback conditions were included in a three-factorial ANOVA, only the main effects for p -changes, $F(2,144) = 53.18$, $\eta^2 = .425$, $p < .001$, and for n -changes, $F(1,72) = 69.17$, $\eta^2 = .490$, $p < .001$, were highly significant. The absence of any interactions (all F s < 1.75) testifies to the ineffectiveness of the debriefing and the feedback treatment. Even full feedback did not reduce the unwarranted influence of n -changes on judgments of p -changes.

Stage 3. The feedback may have taken some time to unfold its influence and therefore may only be visible in the following block. So let us consider the transfer performance on the last trial block, which provided an equivalent task for all three conditions. Figure 3 reveals that all three conditions produced the familiar pattern: strong main effects for p -changes, simple effects $F(2,48) = 12.880$, 24.739 , 52.322 , all $p < .001$, $\eta^2 = .501$, $.315$, $.613$, for the no-feedback, partial-feedback and full-feedback conditions, respectively, but also main effects for n -changes, simple effects $F(1,24) = 20.814$, 11.615 , 13.772 , all $p < .001$, $\eta^2 = .179$, $.219$, $.488$. Weak p -changes $\times$ n -changes interactions reflect that the biasing impact of n -changes is not exactly the same at all levels of p -changes, $F(2,48) = 2.421$, $p = .100$, $\eta^2 = .092$, $F(2,48) = 5.163$, $p = .009$, and $F(2,48) = 3.686$, $p = .032$, $\eta^2 = .133$. Shaded bars are again consistently above the black bars in Figure 3.

Figure 3: Mean change judgments in Stage 3 of Experiment 1 as a function of p -change and n -change and three feedback conditions

Yet, when feedback conditions are included in a three-factorial ANOVA, a weak but significant feedback x p -changes interaction shows up, $F(4,144) = 3.184$, $p = .015$, $\eta^2 = .081$, along with the main effects for p -changes, $F(2,144) = 82.23$, $p < .001$, $\eta^2 = .533$, and n -changes, $F(1,72) = 44.61$, $p < .001$, $\eta^2 = .383$, and the p -changes x n -changes interaction, $F(2,144) = 8.50$, $p < .001$, $\eta^2 = .106$.

Figure 3 shows that the average increase in change judgments from left to right is somewhat enhanced in both feedback conditions. Apparently feedback slightly mitigated the bias but the effect was too weak to undo the strong anomalies. Any learning from debriefing and feedback, or from extended experience across 72 change-judgment trials, was at best very modest.

Experiment 2

Up to now, all experiments were based on change detection between successively presented samples. Participants had to make inferences about changes from a comparison of a current sample and a memory representation of the preceding sample that was no longer visible on screen. This set-up not only places memory demands on the judgment task but also raises the question of whether the information that is encoded in memory includes both p and n . If only the proportion p is encoded but not the sample size n , this might explain the ineffectiveness of the feedback in Experiment 1. Thus, to profit from accuracy feedback and from the debriefing that an increasing n induces a bias toward “Increase” judgments, it is essential to know whether the preceding n was smaller than the present n or not. This condition may not be met when samples are presented successively. Moreover, each sample in the successive set-up enters two different comparisons, with the preceding and the subsequent sample, which may cause interference. We therefore resorted to a simultaneous presentation format, such that each trial involved two new samples presented on the left and right side of the same screen. Anomalies in change detection were already found to extend to this simultaneous format (Fiedler et al., 2016). Proportional increases (decreases) from the left to the right sample on the screen were easily detected when n also increased (decreased) from left to right but were difficult to extract when n changed in the opposite direction. Note also that simultaneous presentation of both samples allows for

a more vivid and more informative feedback. Outcome feedback about whether a judgment was correct or incorrect can be attached to a display of an exact copy of the two samples on the screen.

In this and all subsequent experiments, we modified several aspects of the sampling algorithm and the within-participants design. First, the use of simultaneous sample pairs allowed us to manipulate p -changes and n -changes independently of the n of the preceding trial. Thus, every stimulus sample could be simply generated by sampling $n = 8$ or $n = 16$ elements from a binomial distribution with a given parameter p . Second, all change trials involved p -changes of the easier kind, $|\Delta p| = .20$, thus facilitating the p -assessment task. And last but not least, the sample size n only changed on the 12 increasing- p trials and the 12 decreasing- p trials of a block, but not on the 12 equal- p trials in a block, in which the size of both samples was either $n = 8$ or $n = 16$. Thus, the enhanced n -sensitivity observed in Experiment 1 for equal- p trials (when only n varies) could no longer exaggerate the resulting bias. Another factor that may have complicated p -change assessment, namely, that only small n can be extracted in the given time whereas only focal symbols can be counted for large- n samples, can be checked by comparing the accuracy for small and large n on equal- p trials. The failure to encode large n should reduce the accuracy for p changes in large- n trials. Moreover, the modified design should further facilitate the assessment of p -changes independently of n , and the utilization of feedback information, because the metacognitive corrections required for trials of increasing or decreasing n are only well defined for trials of increasing or decreasing p .

Methods

Participants and Design. Seventy-eight male and female students of Heidelberg University participated either for payment (about 6 € for 45 minutes) or to meet a study requirement. They were randomly allocated to one of two groups (no feedback or full feedback); no partial feedback condition was included. Within participants, p -changes (12 decreasing, 12 equal, 12 increasing) and n -changes (6 decreasing and 6 increasing per level of p -change) varied. Only unequal- p trials involved n -changes.

First sample of white symbols presented at a rate of one symbol per 100 ms in random order and in random locations in a black rectangular frame surrounded by a red-brown screen

Second sample is presented in the same format

Response instruction prompt is added in bright white color on top and in the bottom of the screen

Feedback to participant's response consists of an indication of the correct response key above the sample frame and the population parameters below, highlighted in light green

Figure 4: Four stages of the revised change-judgment task using simultaneously presented samples

Materials and procedures. The entire procedure remained largely the same as in Experiment 1, except for a few minimal modifications in the task instructions that were required to explain the simultaneous presentation format (see Figure 4). Thus, participants were now told that “in each trial of a longer series, you will be shown two samples of binary symbols, ‘≡’ and ‘o’, appearing at varying rates”. Later they were asked, “Were the two samples drawn from the same universe? Or did they originate in different universes, and if so, in which of the universes was the ‘≡’-proportion higher?” The wording of the debriefing had to be adjusted slightly.

The order in which stimulus samples were drawn from the same p and n parameters was newly randomized for each participant and for each of three trial blocks. Response latencies were assessed.

Results and Discussion

Stage 1. Recall that because only unequal- p trials were crossed orthogonally with n -changes, equal- p trials are excluded from the primary analysis of the change judgment scores (mean scores on a scale from -1 to $+1$). The results of the first baseline block of trials, prior to the feedback manipulation, again exhibit the familiar anomalies (Figure 5). Pooling across both feedback groups, an ANOVA of mean change judgment scores yielded two main effects, for p -changes, $F(1,77) = 464.131$, $p < .001$, $\eta^2 = .874$, but also for n -changes, $F(1,77) = 113.178$, $p < .001$, $\eta^2 = .092$. Thus, although judgments were clearly sensitive to relevant changes in p , they were also systematically biased toward irrelevant changes in absolute sample size n . Again, the strength of this bias was immense. A comparison of the shaded and black bars in Figure 5 reveals that the impact of n -changes was asymmetrically stronger for increasing than for decreasing proportions, as manifested in a significant interaction, $F(1,77) = 67.862$, $p < .001$, $\eta^2 = .468$.

Stage 2. The performance during the middle trial block did not diverge from this basic pattern in either condition. Two main effects were obtained both for the no-feedback condition, $F(1,38) =$

198.610, $p < .001$, $\eta^2 = .811$ for p -changes and $F(1,38) = 39.478$, $p < .001$, $\eta^2 = .174$ for n -changes, and for the full-feedback condition, $F(1,38) = 199.738$, $p < .001$, $\eta^2 = .681$ for p -changes and $F(1,38) = 48.786$, $p < .001$, $\eta^2 = .270$ for n -changes. In the former condition, the change-judgment curve (i.e., mean judgments for decreasing, equal, and increasing trials) amounted to -0.842 , -0.077 , 0.030 for decreasing n , as compared to -0.239 , 0.030 , 0.376 for increasing n . In the latter condition with full feedback, the corresponding curves amounted to -0.739 , -0.068 , -0.017 for decreasing n compared to -0.077 , 0.098 , 0.491 for increasing n .

Figure 5: Mean change judgments in Stage 1 of Experiment 2 as a function of p -change and n -change (pooling across feedback conditions)

Stage 3 Transfer. The same ANOVA conducted on change judgments in the transfer task (using different symbols) in the last trial block yielded the same basic pattern: Strong main effects for p -changes both in the no-feedback condition, $F(1,38) = 60.364$, $p < .001$, $\eta^2 = .607$, and in the full-feedback condition, $F(1,38) = 41.560$, $p < .001$, $\eta^2 = .205$, but also substantial main effects for n -changes, $F(1,38) = 34.167$, $p < .001$, $\eta^2 = .073$, and $F(1,38) = 17.327$, $p < .001$, $\eta^2 = .308$. Thus, sensitivity to the relevant proportional changes was accompanied by susceptibility for irrelevant n -

changes. The intrusion of n -changes into judgments of p -changes was similarly strong for the full-feedback condition as for the no-feedback condition (cf. Figure 6), as evident from the weak interaction terms, $F(1,38) = 2.234$, $p = .143$, $\eta^2 = .056$, and $F(1,38) = 4.019$, $p = .052$, $\eta^2 = .096$.

When feedback conditions were included in a three-factorial ANOVA, the main effects for p -changes, $F(1,76) = 459.64$, $p < .001$, $\eta^2 = .872$, and for n -changes, $F(1,76) = 112.36$, $p < .001$, $\eta^2 = .101$, appeared more strongly. A significant interaction between both factors, $F(1,76) = 67.493$, $p < .001$, $\eta^2 = .483$, reflects superior p -sensitivity for increasing (vs. decreasing) n . However, crucially, all interactions involving the feedback factor were negligible, all $F_s < 1$.

Figure 6: Mean change judgments in Stage 3 of Experiment 2 as a function of p -change and n -change and feedback conditions

Thus, the intrusion of n -changes into judgments of p -changes was neither eliminated nor substantially reduced during the second trial block, even though participants were explicitly debriefed and received outcome feedback on a trial-by-trial basis and even though judges were clearly sensitive to p -changes. There was no evidence from equal- p samples for the failure to extract n from large samples, as the accuracy of p -judgments was higher when both samples in a trial were large than when both samples were small, $F(1,77) = 5.54$, $p < .021$, $\eta^2 = .067$ (across all 78 participants). More

generally, despite several procedural changes supposed to facilitate the task in Experiment 2, the same judgment biases persisted. If anything, the simultaneous presentation format served to eliminate the modest feedback effect obtained in Experiment 1 with successively presented samples.

Note on response latencies. Unlike in Experiment 1, the subroutine used in Experiments 2 and 3 for response assessment delivered latency data. Although we did not have any chronometric process hypothesis, it seems worthwhile to briefly summarize these latency results. (a) Across all three stages, individual participants' mean latencies for equal ($M = 1301, 999, 1137$; $SD = 649, 708, 618$) and unequal trials ($M = 1240, 1032, 1000$; $SD = 793, 728, 643$) did not differ significantly (all $t(76) < 1.4$); (b) average within-participants correlations between judgment latency and accuracy were very small but systematically negative (shrinking across stages from $-.063$ to $-.067$ to $-.029$ ($SD = .175, .171, .198$); $t(77) = -3.19, -3.47, -1.29$); and (c) no systematic latency differences were obtained between p -change and n -change conditions. These subsidiary latency findings (a) lend no support to a bias toward or against responding "Equal"; (b) rule out a speed-accuracy tradeoff, and (c) suggest that change judgments depended on online processes (during sample presentation) rather than on counting or difficulty experienced during the response-time window, which ought to be sensitive to trial type.

Experiment 3

As another potentially effective debiasing intervention, besides debriefing, feedback, and simultaneous sample presentation, we picked up the previous finding that p -changes can be detected independently of n -changes when samples are described as normalized percentages rather than experienced inductively as an extension in time and space (cf. Fiedler et al., 2016, Experiment 2). Thus, whereas an increase from 8 focal features out of $n = 16$ to 6 out of $n = 8$ is hard to detect, an increase from "50% in a sample of $n = 16$ " to "75% in a sample of $n = 8$ " is straightforward. Thus, extra trust in large samples is not at work. Change judgments may thus be debiased when judges are asked to translate the presented sample proportions into percentage estimates before they make their change judgments. The intrusion of n into judgments of p changes might be eliminated or diminished if

judges estimate normalized percentages before they make comparative judgments. Experiment 3 thus constitutes an extended replication of Experiment 2 with one additional manipulation between participants, namely, whether or not participants estimated the proportion of the critical symbol before they made their change judgments during the second trial block.

Methods

Participants and design. Ninety-six male and female students of Heidelberg University participated either for payment (6 € for a 45 minutes session) or to meet a study requirement. They were randomly allocated to four experimental groups resulting from two between-participants manipulations: feedback (control vs. full feedback) and percentage estimation (with vs. without additional percentage estimates). The within-participants design remained the same.

Materials and procedures. While the procedure in the no-estimation condition was the same as in Experiment 2, each trial in the estimation condition started with the presentation of two samples using the same format and timing as in Experiment 2. Two seconds after they were completed, the two samples within the two frames were erased and percentage estimates were prompted: “What percentage =? Please estimate”. They entered their estimates for the left and then for the right sample in two input windows centered in the sample frame. The percentage-estimate manipulation was crossed orthogonally with the same feedback manipulation as in Experiment 2.

Results

Stage 1. Let us first consider the baseline performance in the initial trial block, pooling the data from all four groups as all treatments started after the first block. Figure 7 reveals by and large the same pattern as in preceding experiments. A strong main effect for p -changes, $F(1,95) = 445.409$, $p < .001$, $\eta^2 = .698$, came along with a main effect for n -changes, $F(1,95) = 68.935$, $p < .001$, $\eta^2 = .254$. The interaction was negligible, $F(1,95) = 0.664$, $p = .417$, $\eta^2 = .007$.

Figure 7: Mean change judgments in Stage 1 of Experiment 3 as a function of p -change and n -change (pooling across feedback conditions)

Stage 2. Would the p -change main effect be greatly enhanced and the n -change main effect be eliminated or at least reduced when in the middle trial block full feedback and explicit estimations highlight the need to assess p independently of n ? A glance at Figure 8 shows that this was hardly the case. Across experimental groups, the main effects of p -change, $F(1,95) = 477.656$, $p < .001$, $\eta^2 = .504$, n -change, $F(1,95) = 66.691$, $p < .001$, $\eta^2 = .314$, remained as strong as in the first block (compare Figures 7 and 8). No interaction was apparent at aggregate level, $F(1,95) = 2.148$, $p = .146$, $\eta^2 = .022$.

With regard to the treatment effects observed in a four-factorial analysis, indeed, the p -change factor came to interact with the percentage estimation factor, $F(1,92) = 7.520$, $p = .007$, $\eta^2 = .076$, with the feedback factor, $F(1,92) = 8.798$, $p = .004$, $\eta^2 = .087$, and with both estimation and feedback in a three-way interaction, $F(1,92) = 7.283$, $p = .008$, $\eta^2 = .073$. The p -change x feedback x estimation interaction was also significant, $F(1,92) = 5.008$, $p = .028$, $\eta^2 = .052$. Judgments were slightly more sensitive to p -changes and less susceptible to n -changes when feedback was provided, especially feedback combined with estimation (see bottom right chart in Figure 8). Notably, both treatments

exerted at least significant influences in the expected direction, although the most overwhelming result still is the persistence of the unwarranted impact of n -changes on judgments of p -changes.

Note in passing that change judgments (on a $-1, 0, +1$ scale) and pairwise differences of percentage estimates were highly consistent. Despite the inclusion of a third response option (“Equal”), the average individual consistency rates (for different trial types) ranged from .565 and .725).

Figure 8: Mean change judgments in Stage 2 of Experiment 3 as a function of p -change and n -change, feedback and estimation conditions

Stage 3 Transfer. The same disillusioning conclusion holds for the transfer performance in the last trial block. Sensitivity to p -changes was slightly enhanced relative to the unwarranted influence of n -changes (Figure 9). Still, in a four-factorial ANOVA, highly significant main effects were obtained not only for p -changes, $F(1,92) = 459.20$, $p < .001$, $\eta^2 = .833$, but also for n -changes, $F(1,92) = 39.56$, $p < .001$, $\eta^2 = .301$, but no p -changes $\times$ n -changes interaction, $F(1,92) = 0.086$, $p = .770$, $\eta^2 = .001$.

Figure 9: Mean change judgments in Stage 3 of Experiment 3 as a function of p -change and n -change, feedback and estimation conditions

Feedback and percentage estimation continued to exert modest but significant moderating influences (see Figure 9). Two higher-order interactions were observed, for p -changes x feedback x percentage estimation, $F(1,92) = 19.38, p < .001, \eta^2 = .174$, and for p -changes x n -changes x feedback x percentage estimation, $F(1,92) = 4.221, p = .043, \eta^2 = .044$, reflecting somewhat enhanced sensitivity to p -changes relative to n -changes in the feedback plus percentage-estimation condition.⁵

We refrain from documenting latency results, which strongly resembled those of Experiment 2.

Interestingly, then, providing explicit percentage estimates before change judgments, along with full feedback, did indeed slightly reduce the unwarranted influence of n -changes on judgments of p -changes. From a metacognitive perspective, this means that spontaneous percentage estimation can help to improve change-detection performance. However, the impact of n -changes on judgments of p -change was still substantial and continued to impair change judgments in all feedback and estimation conditions. The persistence of MM remained the most overwhelming empirical result.

Discussion

To summarize, p -sensitivity is slightly enhanced after estimation, during feedback, and when estimation and feedback coincide. However, again, although both treatments seem to exert to some degree the expected (local) influence on the immediate training set, the unwarranted impact of n -changes remains highly significant in all four experimental groups.

After three experiments with nine conditions, there was little evidence for effective debiasing in change judgments, even after open debriefing, trial-by-trial feedback, explicit estimation tasks, and several blocks of learning experience. On the one hand, these persistent anomalies in the detection of change are not too surprising from our metacognitive perspective. Because n is typically unknown, a generalized summation algorithm may be at work distorting the retrograde application of an averaging

⁵ We refrain from interpreting a two-way feedback x estimation interaction, $F(1,92) = 6.066, \eta^2 = .062, p = .016$; judgments were generally higher when feedback and estimation tasks were either jointly present or jointly absent.

rule when the end of a sample finally reveals n . On the other hand, however, it is possible that we simply failed to create the very conditions under which judges easily manage to separate p from n .

This possibility is indeed counted on by several colleagues who may not disagree with the notion of MM but nevertheless believe that learning to separate p from n is just a matter of appropriate training. To deal with this possibility, and to rule out a shortsighted mistake in our research set-up, we here started a collective research project, inviting several interested colleagues in the area of metacognition and decision making to help us optimize a debiasing treatment. The remainder of this article is devoted to this collective project.

Experiment 4

The instruction text used to solicit all collaborators' advice is given in Appendix A. To summarize, we prompted suggestions to optimize the procedure and the format (timing, wording, location on screen) of the feedback; the verbal debiasing instructions; the embedding cover story; presentation style of stimulus samples; performance-contingent payoff; and any other spontaneously offered advice. The suggestions provided in response to these prompts can be categorized into three classes: *Social debiasing*: With reference to earlier work by Koriat (2008) and Sperber, Clément, Heintz, Mascaro, Mercier, Origgi, and Wilson (2010), it was suggested that collective work in dyads or playing interpersonal games should help to overcome the bias. While we are enthusiastic about such social treatments, this idea calls for qualitatively different experiments that we postpone to the future. *Incentives and motivating task settings*: Accuracy incentives should be provided in a more effective manner (cf. Camerer & Hogarth, 1999) and the judgment task should be reframed competitively, as a betting game. The idea here was that proportions matter when it comes down to making consequential choices between lotteries. *Timing and feedback format*: The debiasing treatment should rely on more comprehensible debriefing instructions and more effective feedback procedures. Feedback should not only provide information about what judgments are “correct” versus “incorrect”. It is also essential to relate this feedback (a) to an effectively coded representation of the stimulus samples (without time

pressure) and (b) to explicit instructions about how to compute proportions properly. Based on these collectively generated ideas and on the experience gathered in Experiments 1 to 3, we designed the following experiment that incorporates the greatest part of all collaborators' convergent suggestions.

Keeping within the same general task setting as in all previous studies, the collectively designed Experiment 4 accorded to the following features: (1) The between-participants design only consisted of a single factor, contrasting the performance of an experimental group exposed to debiasing training and a control group without this treatment. (2) Three trial blocks were included within participants. The debiasing manipulation was administered in the first trial block. Whereas one condition underwent an intensive debiasing training, the other (control) group was exposed to the standard change-detection task, to which both conditions were exposed in the second trial block. Finally, the last block involved a slightly modified change-judgment task, combined with a betting game that highlighted the need to consider proportions in the urns.

As detailed in the method section below, the debiasing training in the first block combined several elements: (a) an optimized cover story and instruction text referring to the “balls and urns” metaphor, which should help to clarify the distinction of the sample and the latent universe (i.e., the urn); (b) explicit debriefing instructions; (c) numerical proportion estimates that had ameliorated the anomalies in Experiment 3; (d) a more elaborate feedback procedure, linking correctness to an exact copy of the juxtaposed stimulus samples; and (e) the possibility to correct for the initial judgment in a second judgment provided under optimally facilitative conditions.

It could be expected that all these features together should clarify the change detection task and the feedback. Still, granting our metacognitive account of the difficulty to normalize a growing sample p for an unknown n , the anomalies may persist even under such highly auspicious conditions.

Methods

Participants and Design. One hundred and seventy male and female students of the University of Heidelberg participated in Experiment 4. They were drawn from the same pool as in all prior

experiments. They were randomly allocated to the debiasing or the control condition. Within participants, the same p -change $\times$ n -change design was used as in Experiments 2 and 3.

Materials and procedures. The introduction to the detection of change task was modified slightly. To facilitate the stimulus encoding, the two symbols (“≡” and “o”, as before) were presented in yellow and blue color, respectively. Drawing on the familiar “balls and urns” metaphor, each sample was said to stem from an urn rather than from a universe. The resulting core instruction now read: “Which one of the two samples was drawn from the urn with the higher proportion of yellow symbols? It is also possible that both samples stem from urns with the same yellow proportion.” The number of warm-up trials was increased from two to eight.

Block 1. Apart from verbal instructions, the first block of 36 trials in the no-debiasing control condition was exactly the same as in preceding experiments, offering a chance to replicate the canonical pattern. Participants in the control condition only read, “So these were the warm-up trials ... Relax a moment and then start (using any key press) a second block of trials involving the same task.”

In the debiasing condition, the first block comprised only 24 trials, to compensate for the much more elaborate style and for the longer duration of training trials. In this condition, the instruction was extended as follows:

“Before we make a start, however, you will first receive a special training that should help you to respond optimally and to avoid mistakes.

As it turned out, such proportional judgments are erroneously often more sensitive to the total number of symbols than to the relative proportion of yellow rather than blue symbols.

What really counts logically is the ratio of yellow symbols to the total number of all symbols in a sample. When I have, say, a sample of 6 yellow and 2 blue symbols on the left this amounts to 75% yellow symbols. If I now have 8 yellow and 8 blue symbols on the right, then there are only 50% yellow symbols.

This means: Even though there are more yellow symbols (8) on the right than on the left (6), the ratio on the right is nevertheless only 50% compared to 75% on the left. What counts is indeed the ratio of yellow versus blue symbols, not the size of the sample.

To estimate the proportion accurately, you therefore have to attend to the number of yellow symbols as well as the number of all symbols.”

On the next instruction screen, they were then prepared for the percentage estimation task and for the feedback training:

“You are now supposed to undergo the following training of the critical distinction between the number of yellow symbols and the size of the sample.

As shown before, a series of training tasks will each involve two samples of yellow and blue symbols. A moment later, the samples will disappear again and you are asked to type into two black boxes what % of yellow symbols was included in both samples.

Thereafter, an IDENTICAL copy of both samples will reappear, but this time they will not be erased but remain onscreen, when you are again asked to estimate the percentage of yellow symbols. This repetition under highly facilitative conditions should enable you to correct for erroneous estimates, so that proportions are clearly distinguished from sample sizes.

Finally – while both samples are still visible – you are supposed to decide whether the left or right sample is drawn from an urn with a larger proportion of yellow signs or from an urn with an equal yellow-proportion. Furthermore, you will receive feedback indicating if your decision was right or wrong. So during this training phase, you are given an opportunity to test out and to improve your proportional judgments.”

The feedback format was further designed to maximize the chance to correct for erroneous change judgments. Whether responses were “Correct” or “Wrong” was indicated in large green or red letters, respectively, on top, while exact copies of the two samples were visible and the correct

percentage of $\equiv$ symbols was indicated underneath each frame. Feedback also included explicit information about which arrow key ($\leftarrow$, $\rightarrow$, or $\downarrow$) would have been correct.

Block 2. The second block was identical to the first block in the control condition, involving the same 36 pairs of n and p , parameters, presented in a new random order.

Block 3. The last trial block again served as a transfer test, but the normal change judgment task was modified as a betting game with performance-contingent payment, which was intended to further enhance the motivation to attend to proportions proper:

“In the last part of the study, the rules of the game will be slightly modified.

Imagine the urns from which the samples are drawn are actually lotteries, in which you can actually participate. Your winning chances depend on the proportion of yellow symbols. You will win whenever you bet on the lottery with the higher proportion of yellow balls (symbols).”

After the instruction to use the same arrow response keys as before to choose between lotteries, it was then explained that a black response box would be opened above the frame of the chosen lottery and that they could type in how many points they wanted to bet. They were told: “You can bet between 1 and 10 points. If you choose correctly, you will win this amount. If you choose incorrectly, this amount will be subtracted from your account. If you don’t make a choice, your account will remain unchanged.” Thus, on every trial of the last block the same initial change-assessment response was followed by a betting game referring to the same sample pair.

Note that this betting game in the last trial block should motivate and facilitate the assessment of proportions and effectively prevent judges from confusing proportions with absolute numbers. After the last trial block, participants were thanked, debriefed, and paid.

Results and Discussion

Stage 1. During the first block, the control condition yielded the typical pattern (Figure 10, top left panel): an overwhelming main effect for p -changes, $F(1,88) = 722.29$, $p < .001$, $\eta^2 = .891$, together

with another strong main effect for the unwarranted influence of n -changes, $F(1,88) = 40.53, p < .001$, $\eta^2 = .315$. The interaction was negligible, $F(1,88) = 0.40, p = .529, \eta^2 = .005$.

Figure 10: Mean change judgments in Experiment 4 as a function of p -change and n -change, stages, and debiasing condition versus control condition

In the debiasing condition (top right panel), the facilitative instructions and unlimited presentation time enabled judges to concentrate on the task-relevant aspect. As a consequence, change judgments (in the 2nd run without time limit) only exhibited a p -change main effect, $F(1,80) = 624.17$, $p < .001$, $\eta^2 = .886$, but no n -change main effect, $F(1,80) = 2.81$, $p = .098$, $\eta^2 = .034$, and no interaction, $F(1,80) = 2.17$, $p = .145$, $\eta^2 = .026$.

Not surprisingly, a comparison of both conditions revealed a significant difference in sensitivity to n -changes (mean difference of change judgments for increasing- n minus decreasing- n trials) was clearly significant, $t(168) = 3.958$, $p < .001$, $d = 0.609$. In the control condition, 64 of all 89 participants provided on average higher change-judgments on increasing- n than decreasing- n trials; in the debiasing condition, this was only the case for 47 out of 81 participants, $\chi(1) = 3.61$, $p = .057$. Although performance improved as a consequence of the decision aids (double presentation; double estimation; feedback etc.), other measures indicate that n -changes continued to intrude into p -change assessment even under auspicious conditions.

Figure 11: Mean percentage estimates as a function of p -change and n -change and estimation stages during the debiasing treatment of Experiment 4

Percentage estimates. The manipulation of n -changes biased the percentage estimates of p changes. This is clearly evident in the first estimates provided after the initial brief presentation (left chart in Figure 11), which produced not only a strong main effect for actual changes in p , $F(1,79) = 354.79$, $p < .001$, $\eta^2 = .818$, but also for n -changes, $F(1,79) = 22.94$, $p < .001$, $\eta^2 = .225$. Remarkably, the second percentage estimates (right chart) provided after unlimited exposure also reflected both main effects, not only for p -changes, $F(1,79) = 553.02$, $p < .001$, $\eta^2 = .875$, but also (though much weaker) for n -changes, $F(1,79) = 6.04$, $p = .016$, $\eta^2 = .071$.⁶

Although percentage estimates were somewhat regressive (i.e., less extreme than actually presented percentage differences of -20% and $+20\%$, respectively), they were rather accurate. Accuracy increased as expected from the first to the second estimates under extremely facilitative conditions. Thus, in spite of the ease of the transparent estimation task, together with the explicit debriefing, the permanent trial-by-trial feedback, and the explicit instruction not to fall prey to misleading sample-size effects, even the second estimate of sample percentages in the debiasing condition continued to reflect a biasing impact of sample size.

Stage 2. How about the performance during Stage 2, once debiasing training was over? An analysis of change judgments shows that in the control condition (Figure 10, middle left panel) the p -changes main effect, $F(1,88) = 795.81$, $p < .001$, $\eta^2 = .900$, and the n -changes main effect, $F(1,88) = 18.32$, $p < .001$, $\eta^2 = .172$, were still visible, as in all previous experiments, but no interaction, $F(1,88) = 0.85$, $p = .359$, $\eta^2 = .010$. In the debiasing condition (Figure 10, middle right panel), the acquired ability to ignore n -changes was still noticeable: Only the p -changes main effect was obtained, $F(1,80) = 722.45$, $p < .001$, $\eta^2 = .900$, whereas the n -changes main effect fell short of significance, $F(1,80) =$

⁶ The p -changes x n -changes interaction was negligible for first estimates, $F(1,79) = 0.92$, and non-significant for second estimates, $F(1,79) = 3.47$, $p = .066$; n -size effects were slightly stronger for decreasing than for increasing p .

1.79, $p = .185$, $\eta^2 = .022$, and no interaction was obtained, $F(1,80) = 0.58$, $p = .449$, $\eta^2 = .007$. A comparison of control and debiasing conditions showed that the tendency to provide higher change judgments for increasing n than for decreasing n was significantly higher in the control condition ($D = -0.254$, $SD = 0.559$) than in the debiasing condition, ($D = -.062$, $SD = 0.415$), $t(168) = -2.521$, $p < .013$, $d = -0.388$.⁷

This suggests that the new ability to assess p -changes separately from n -changes was not contingent on continued feedback. The lesson gained from the debiasing training was sufficient to maintain the improved performance during the next stage, which involved the same stimulus format and the same Δp distribution but without any online feedback and no reminder of the debriefing. Thus, for the first time in an extended series of experiments it appears that it was possible at least to reduce the misleading impact of n -changes on the assessment of p -changes.

Stage 3 transfer. However, an analysis of the transfer performance in the final trial block (with new symbols and identical treatment of both experimental groups) suggests that the debiasing-training effect was only transitory and largely restricted to the training set; it did not result in a transferrable stable learning effect. As apparent from a comparison of shaded and filled bars in Figure 10 (bottom charts), both conditions not only exhibit an overwhelming p -change main effect, $F(1,88) = 813.96$, $p < .001$, $\eta^2 = .902$, and $F(1,80) = 489.61$, $p < .001$, $\eta^2 = .860$, for the control and debiasing conditions, respectively. Rather, the unwarranted n -changes main effect was obtained not only in the control condition, $F(1,88) = 11.52$, $p = .001$, $\eta^2 = .116$, but also at equal strength in the debiasing condition, $F(1,80) = 13.68$, $p < .001$, $\eta^2 = .146$. Sensitivity to unwarranted samples-size effects no longer differed between control ($D = -0.180$, $SD = 0.500$) and debiasing conditions ($D = -0.160$, $SD = 0.390$), $t(168) = -0.281$, $p < .013$, $d = -0.043$. No interactions were obtained between training and within-participants factors (all F s < 1.55), as evident from highly similar patterns of the two bottom charts of Figure 10.

⁷ The discrimination of increasing and decreasing p did not differ, $t(168) = 0.051$.

Betting game results. On every trial in the final block, participants not only provided a change judgment but they were also engaged in a betting game. They could bet between 1 and 10 points on the sample they supposed to stem from a population with a higher p . Their subsequent payment was contingent on these betting decisions. The question of interest was whether such a betting game would motivate or force participants to better understand the logic of the change-detection task.

Granting that actual p -changes were held constant at $|\Delta p| = .20$ across all unequal trials, the null hypothesis was that a rational decision maker should bet on average the same amount, regardless of whether p decreased or increased and regardless of whether n decreased or increased. The typical bias obtained in change judgments, in contrast, should be manifested in a p -change $\times$ n -change interaction; judges should bet more on p -“increase” (p -“decrease”) judgments when n also increased (decreased).

Figure 12: Mean betting decisions in Experiment 4 as a function of p -change and n -change and control condition versus debiasing conditions

Not surprisingly, in the control condition the betting data did exhibit such an interaction, $F(1,88) = 18.32, p < .001, \eta^2 = .172$. This suggests that the incentive value of the betting game alone, or any latent learning improvement reached after 2 times 36 trials in the two preceding trial blocks, were not

sufficient to erase the unwarranted impact of n -changes on betting decisions in the final trial block.

Figure 12 shows that the interaction was somewhat reduced in the debiasing condition, but it was also significant $F(1,80) = 5.260, p = .024, \eta^2 = .062$. All interactions involving the conditions factor (with p -change, n -change or both) were negligible (all F s < 1.50).

General Discussion

Our collective multi-lab approach to defeating an obstinate anomaly in inductive reasoning converged in a stable pattern of empirical results. When assessing changing proportions of a focal outcome from limited samples of binary observations, participants were highly sensitive to the actual p in the universe, from which samples were drawn. Indeed, the main effect for p -sensitivity was the strongest finding obtained across all experiments, ruling out the possibility that participants were simply under-motivated or that they may have misunderstood the experimental instructions, counting only the frequency of focal outcomes and ignoring the ratio of focal to non-focal outcomes.⁸ Rather, our participants were quite accurate at inferring changes in latent populations from noisy stimulus samples, the manifest proportions of which could vary considerably around the true p .

However, despite this basic accuracy motivation and sensitivity to p , as required by the task instructions, there was ample evidence for a cognitive bias: A strong and significant main effect for n -changes, obtained consistently across experiments and task conditions, testifies to the impact of metacognitive myopia. Participants were unable to assess p , the relative proportion of a focal outcome, independently of n , the absolute size of the sample. Increasing (decreasing) n was mistaken as increases (decreases) in p . Increasing n facilitated the recognition of increasing p , but interfered with the assessment of decreasing p . Conversely, decreasing n facilitated the correct assessment of decreasing p but reduced correct responding on increasing- p trials.

⁸ Evidence against such a superficial counting strategy comes from an experiment reported by Fiedler et al. (2016). Had participants only counted focal outcomes, the same p -changes would have appeared stronger in larger samples.

While the reported analyses are all based on change judgments (coding -1 , 0 , and $+1$ for “decrease”, “unchanged” and “increase judgments, respectively), equivalent findings were obtained in analyses of correctness proportions. Here the canonical pattern is a marked disordinal interaction between p -changes and n -changes: Correctness rates are high when both p and n change in the same direction but low when p and n change in opposite directions. We refrain from reporting these findings, which follow logically from the reported change judgments.

Crucially, the present findings go way beyond replicating the co-existence of adaptive sensitivity to p -changes and unwarranted sensitivity to n -changes that was already demonstrated in previous research with successively presented samples (Fiedler et al., 2016). Evidently, the full pattern of results was replicated to the same degree when simultaneously presented samples ruled out memory demands and when the elimination of no-change trials reduced the complexity and the indeterminacy of the task. Thus, including only upward changes of $\Delta p = +.2$ or downward changes of $\Delta p = -.2$ ruled out (in Experiments 2 to 4) the potentially confusing case of minor sample changes obscuring latent invariance ($\Delta p = 0$), as in Experiment 1 and in previous research. Moreover, the biasing influence of n changes was neither eliminated by explicit debriefing instructions nor by immediate trial-by-trial feedback of the correct change status presented with a perfect copy of the simultaneously presented stimulus samples (in Experiments 2 and 3). Together with the debriefing about the danger to mistake increasing (decreasing) n for an increasing (decreasing) p , such comprehensive feedback must have underscored the biasing influence of n changes. The anomalies also persisted (in Experiment 3) when participants first estimated the observed percentage of the focal feature in the stimulus samples, although the slightly reduced strength of bias in the percentage-estimation condition corroborates previous evidence that a normalized percentage format can be helpful.

Finally, an elaborate debiasing training was developed in Experiment 4 in collaboration between ten researchers working in different labs and holding different expectations about the malleability of

metacognitive myopia. While all collaboration partners were clearly informed about metacognitive myopia in general and about the specific confusion of p and n changes in particular, most of them believed that it should be possible to eliminate the intrusion of n -changes into judgments of p changes: through careful instructions, using the balls-and-urns analogy to explain sample-based inferences, unlimited sample presentation times during the debiasing training, sufficient time to digest the feedback, interspersed percentage estimates, using a betting game to mobilize latent metacognitive insights, and through combination of all these measures. The purpose here was to exploit collective expertise in an attempt to devise an effective debiasing training.

The jointly developed debiasing interventions were obviously attended to, leading to improved change judgments during the training phase, as long as the stimulus presentation time was unlimited and the feedback was provided transparently and blatantly. This positive evidence for a local debiasing effect is not trivial; it testifies to the participants' motivation and ability to understand the instructions in all respects. Thus, notably, during the debiasing block, the confusion of n changes and p changes was almost (though not completely) eliminated.

However, this apparent success was only local. There was little sign of transfer to a final stage of change judgments, when the influence of the debiasing training was tested under "normal" conditions, in the absence of feedback, debriefing, percentage estimates, and the other simplifying conditions, using novel stimuli for the change-detection tasks. In this crucial transfer test, no sustainable learning effect was apparent. The old anomalies were visible to the same extent in the debiasing condition as in the control condition. Moreover, the incentives of a betting game did not prevent participants from confusing n and p changes. It seems fair to conclude that – despite some local success in overcoming a strong bias under most auspicious conditions – the provocative bias in change judgment persisted until the end of a long series of debiasing attempts.

This may of course not be the last word. Successful debiasing may take more time and maybe more imagination in devising even more efficient remedies than the present ones. Maybe it is essential

to not only explain and illustrate the anomaly through debriefing and feedback but also to exercise the correction process for a sufficiently long time. Maybe social motivation and epistemic vigilance can be evoked to mobilize better monitoring and control performance (Mata, Fiedler, Ferreira & Almeida, 2013; Sperber, Clément, Heintz, Mascaro, Mercier, Origgi & Wilson, 2010). We have seen that self-generated percentage estimates may be a useful tool that could be exploited. It is also possible that implicit measures of change detection, as assessed by eye-tracking or priming effects, are more likely to separate p from n than explicit judgments.

However, all these optimistic considerations must not distract us from noting the strength and persistence of metacognitive myopia observed across the present experiments. But what can we learn from these findings? What does the impact of n changes on detection of p changes tell us about the role of metacognition for change assessment and inductive reasoning?

At the most task-specific level, first of all, there are good reasons why metacognitive myopia is so tenacious in the context of this particular task of assessing binary samples. A very basic law in psychology is that learning increases with the number of learning trials or observations. We cannot tell our cognitive system to stop learning from repeated observations. Just as one cannot tell an organism to stop learning from pairings of conditional and unconditional stimuli in a conditioning experiment, we cannot decide to cease learning from further observations in a binary sampling task. Binomial sampling may be not so different from conditioning. Imagine an Internet search for positive (+) versus negative (−) hotel evaluations. Let us assume that hotel A receives five positive and only one slightly negative review, creating a rather positive impression. Now assume that another hotel B receives the same results, which are then however repeated in another set of five positive plus one slightly negative review. Why should added observations not create an increment in positive evaluation, just as six new trials in a conditioning effect result in stronger evaluative learning?

To explain why inductive-statistical learning depends on n , it is instructive to consider the constraints imposed on the learning process. As explained at the outset, a binary learning or induction

process must be sensitive to sample size although n is not known beforehand. For an n -independent encoding of 5+ and 1– observations, each elementary observation must be given a weight of $1/6$. In general, each coded elementary observation must be weighted by $1/n$. However, if n is unknown, what increment and decrement should be coded for each added + and –, respectively? It is hardly possible to start coding $1/6$ increments and then, when the sample happens to be extended beyond, to correct the coded value retrospectively. Although Bayesian calculus principally allows for such a correction, it is unrealistic because it would be necessary to keep track of sizes of all ongoing sampling tasks, let alone the unrealistic mental calculation load. A much more realistic assumption is that human observers either resort to summation (rather than averaging) from the beginning, ignoring n and only assessing the cardinal frequencies $f(+)$ and $f(-)$, or attempt to assess proportions based on some default n . This could be the average n of all samples encountered in a sequential experiment. Or, in a context-sensitive process, it could be the n of the previous or neighboring sample. When the previous (or average) n is small (large) the incremental weight given to each observation of a current larger (smaller) sample will be too high (low). This might exactly explain the dependence of p -assessment on n -changes.

Such a view on metacognitive myopia in terms of the mental operations required for binomial encoding suggests that our participants' behavior may be highly adaptive. It may, in fact, reflect the best encoding strategy that a well-adapted organism could apply. From this perspective, the present findings have little to do with irrational behavior. Instead, these results could motivate a novel theoretical approach to understanding the manner in which experienced quantities are represented in memory. Apparently, the mental representation of the same ratio of, say, $r = 5/6 = 10/12 = 0.833$ can look quite different, depending on the encoding history of the underlying inductive inference process. The dependence of p -changes on n -changes may reflect an extensional code that is sensitive to the experienced binomial pattern in the mental representation of proportional quantities. This assumption is reminiscent of the prominent literature on the gap between decisions from experience and from description (Hertwig, Hogarth & Lejarraga, 2018; Wulff, Mergenthaler-Canseco & Hertwig, 2018).

However, although these considerations of logical process constraints may help to understand the origin of the myopia for the confusion of p and n , they do not constitute an acquittal of irrationality at the functional level. After all, under specific conditions, the n -independence of p -judgments can be a logical necessity, and violations of this necessity can be termed irrational. If a teacher's grading decisions are supposed to reflect different students' proportions of correct responses, or their average correctness, detached from the absolute number of responses, then it is irrational and unfair to grade 40+ and 8– higher than 20+ and 4–. But metacognitive myopia is even more tenacious. Even when the same distribution of, say, 6 positive and 2 negative stimuli is simply repeated, and judges know that it is a plain repetition, providing no independent information, and even when they are warned not to be influenced by redundant repetitions, they continue to “learn from repetition” (Fiedler et al., 2018; Unkelbach et al., 2007). One can hardly deny the irrational character of such myopia.

Considering the mechanisms underlying MM from a metacognitive perspective highlights the challenge involved in amending metacognitive monitoring processes and the following control decisions. Attempts to overcome monitoring biases are often challenging. For instance, Castel (2008) attempted to draw participants' attention to serial order effects when memorizing a list of words. Only when the ordinal number appeared without the word to be remembered it was reflected in judgments of learning. Similarly, Yan, Bjork, and Bjork (2016) attempted to guide participants by both experience-based and theory-based debiasing techniques to appreciate the learning benefits of interleaving over blocking their repeated learning cycles. They failed to achieve this goal in five experiments. The sixth experiment succeeded by letting participants experience both study schedules separately in addition to experience- and theory-based debiasing. This combination led a majority of learners to appreciate interleaving advantage over blocking repeated learning. The authors concluded that sense of fluency, pre-existing beliefs, and a strong belief that one is unique in his/her study approach underlie the persistent bias towards blocking over interleaving learning. In meta-reasoning, rather than learning contexts, debiasing attempts often involve encouraging cognitive reflection among participant who do

not perform it spontaneously (see Ackerman & Thompson, 2017 for a review). In this context, if the initial solution (or decision) that comes to mind is accompanied by high Feeling of Rightness, then people provide it with little further consideration. When the Feeling of Rightness regarding the initial solution is lower, then people deliberate and might change their response.

However, even when people acknowledge their lower chance of success, by lower metacognitive judgments, answer changes are relatively rare even after rethinking attempts (e.g., Johnson E. D., Tubau, & De Neys, 2016; Thompson & Johnson S. C., 2014). In the present study, performance was the measure for overcoming the MM. Future studies are called for to consider initial versus later reasoning phases by collecting Feeling of Rightness, response times before and after the initial response that comes to mind, answer changes, and final confidence (see Thompson & Johnson, 2014, as an example of this paradigm), for shedding more light on the underlying metacognitive processes involved in MM. Moreover, metacognitive methodologies may employ factors that affect the extent of the bias at the participant level, task level, and for each particular item (see Ackerman, 2019).

A final remark is in order regarding the goal of our collective project to enrich the current debate how to improve the quality of science in general and to exploit the surplus value of multi-lab research in particular. While the greatest part of this debate revolves around data sharing and exact replication in different places, we believe that other variants of cooperative and competitive collaborations may turn out to be very fruitful and inspiring. Here we wanted to demonstrate how one of the biggest problems in behavioral science – the operationalization of an experimental treatment or intervention – can be profitably tackled in a friendly and at the same time adversarial collaboration.

Appendix A

Question prompts used to prompt the collaborators' advice in the collective project

Granting that the feedback and debiasing instructions we have used so far was not effective enough or that a suboptimal task setting might have not supported the feedback treatment sufficiently, how should the task setting and the feedback procedure be operationalized to eliminate the intrusion of n changes into judgments of p changes? Specifically:

- (1) What format (timing, wording, format and location on the screen) can be expected to optimize the likelihood of feedback learning?
- (2) How would you frame and formulate the verbal debiasing instruction (i.e., the debriefing) supposed to motivate the participants' deliberate attempt to avoid confusing n with p ?
- (3) Within the depicted task setting (i.e., two binary samples presented simultaneously on the left and right half of the screen), how should the two samples be presented optimally, to render the feedback most effective? – Answers to this question might refer to the timing and ordering of the elementary observations from both samples, the nature of the symbols used for the binary outcomes, the embedding task or cover story, performance-enhancing incentives etc.)
- (4) Last but not least, do you have any other suggestions, independent of those solicited by questions (1) to (3), that you believe might improve a debiasing treatment in our paradigm?

References

- Ackerman, R. (2019). Heuristic Cues for Meta-Reasoning Judgments: Review and Methodology. *Psihologijske teme*, 28(1), 1-20.
- Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21, 607-617.
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1-3), 7-42.
- Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, 36(2), 429-437.
- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66(5), 819-829. doi:10.1037/0022-3514.66.5.819
- De Neys, W., & Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences*, 17, 172-178.
- Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, 107(4), 659-676.
- Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and 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.

- Fiedler, K., Hütter, M., Schott, M., & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*.
<https://doi.org/10.1002/bdm.2109>
- Fiedler, K., Kareev, Y., Avrahami, J., Beier, S., Kutzner, F., & Hütter, M. (2016). Anomalies in the detection of change: When changes in sample size are mistaken for changes in proportions. *Memory & Cognition*, *44*(1), 143-161.
- Fiedler, K., Wöllert, F., Tauber, B., & Heß, P. (2013). Applying sampling theories to attitude learning in a virtual school class environment. *Organizational Behavior and Human Decision Processes*, *122*(2), 222-231. doi:10.1016/j.obhdp.2013.08.001
- Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986.
doi:10.1177/1745691617702496
- Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and description: Exploring two paths to knowledge. *Current Directions in Psychological Science*, *27*(2), 123–128.
<https://doi.org/10.1177/0963721417740645>
- Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538-540. doi:10.1097/00001888-199805000-00024
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56-64.
- Johnson-Laird, P. N., Miller, G. A., Galanter, E., Pribram, K. H., Nelson, T. O., & Narens, L. (1992). Theory for metacognition. In T. O. Nelson (Ed.), *Metacognition: Core readings*. (pp. 101–130). Needham Heights, MA: Allyn & Bacon.
- Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*(4), 534-544. doi:10.1037/0022-3514.63.4.534

- Koriat, A. (2008). Alleviating inflation of conditional predictions. *Organizational Behavior and Human Decision Processes*, 106(1), 61-76.
- Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, 105, 353-373.
- Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, 49(3), 486-491.
- Mikulincer, M. (2013). *Human learned helplessness: A coping perspective*. Springer Science & Business Media.
- Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, 51(2), 102-116.
- Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, 106(1), 13-26.
doi:10.1016/j.cognition.2006.12.008
- Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, 143(5), 2000-2019. doi:10.1037/xge0000013
- Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological Science*, 28(10), 1432-1442.
<https://doi.org/10.1177/0956797617711291>
- Price, P. C. (2001). A group size effect on personal risk judgments: Implications for unrealistic optimism. *Memory & Cognition*, 29(4), 578-586. <https://doi.org/10.3758/BF03200459>
- Price, P. C., Smith, A. R., & Lench, H. C. (2006). The effect of target group size on risk judgments and comparative optimism: The more, the riskier. *Journal of Personality and Social Psychology*, 90(3), 382-398. <https://doi.org/10.1037/0022-3514.90.3.382>

- Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18(1), 89-107.
doi:10.1016/j.lindif.2007.03.011
- Smith, A. R., & Price, P. C. (2010). Sample size bias in the estimation of means. *Psychonomic Bulletin & Review*, 17(4), 499-503. doi:10.3758/PBR.17.4.499
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359-393.
- Stanovich, K. (2011). *Rationality and the reflective mind*. Oxford University Press.
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20(2), 215-244.
- Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3), 107-140.
- 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, 37-52.
- Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144(2), 140–176.
<https://doi.org/10.1037/bul0000115>
- Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, 145(7), 918-933.