January 2020

Reference:
Ackerman, R., Bernstein, D. M., & Kumar, R. (in press). Metacognitive hindsight bias. *Memory & Cognition.*

Metacognitive Hindsight Bias

Rakefet Ackerman

Technion—Israel Institute of Technology, Haifa, Israel

Daniel M. Bernstein

Kwantlen Polytechnic University, Surrey, British Columbia, Canada

Ragav Kumar

University of Victoria, Victoria, Canada

Corresponding Author: Rakefet Ackerman, PhD.
Faculty of Industrial Engineering and Management
Technion, Technion City, Haifa 3200003, Israel.
E-mail: ackerman@ie.technion.ac.il

Abstract

Hindsight Bias (HB) is the tendency to see known information as obvious. We studied Metacognitive Hindsight Bias (MC-HB)—a shift away from one's original confidence regarding answers provided before learning the actual facts. In two experiments, participants answered general-knowledge questions in social scenarios and provided their confidence in each answer. Subsequently, they learned answers to half the questions and then recalled their initial answers and confidence. Finally, they re-answered, as a learning check. We measured confidence accuracy by calibration (over/underconfidence) and resolution (discrimination between incorrect and correct answers), expecting them to improve in hindsight. In both experiments, participants displayed robust HB and MC-HB for resolution—improved discrimination between correct and incorrect answers—despite attempts to recall the initial confidence in one's answer. In Experiment 2, promising anonymity to participants eliminated MC-HB, while social scenarios produced MC-HB for both resolution and calibration—indicative of overconfidence. Overall, our findings highlight that in social contexts, recall of confidence in hindsight is more consistent with answers' accuracy than confidence initially was. Social scenarios affect differently HB and MC-HB, thus dissociating these two biases.

Keywords: Memory, Hindsight bias, Metacognition, Confidence, Social Cognition

Hindsight bias (HB) is a robust phenomenon that makes the past seem more predictable than it was. For instance, it is hard for a police investigator to ignore her prior knowledge regarding a criminal event when soliciting a suspect's point of view. Similarly, it is hard for an instructor to ignore his knowledge to understand the source of students' difficulty (see Dror, Morgan, Rando, & Nakhaeizadeh, 2017; Louie, Rajan, & Sibley, 2007).

Many HB studies use a memory design in which participants answer general-knowledge questions (e.g., "In what year did Obama become US President?" – participant answers "2006"). Next, participants learn the correct answers to some questions (e.g., "Obama became President in 2009"), and then try to recall their initial answers to all questions. HB emerges when participants "recall" an answer closer to the correct answer than they initially provided (e.g., 2007). This phenomenon occurs throughout the world, across the lifespan, and in various real-life contexts, including forensic, legal, medical, and investment decisions (see Bernstein, Aßfalg, Kumar, & Ackerman, 2016; Roese & Vohs, 2012).

Of particular relevance to the present study is the finding that HB relates to the uncertainty which accompanied the initial answers. People generally show overconfidence regarding their biased responses in hindsight (Fischhoff, 1975; Hawkins & Hastie, 1990) and tend to show more HB as their confidence in their initial answer is lower (that is, feeling of uncertainty regarding one's knowledge, Pohl & Erdfelder, 2017). When considering interpersonal expertise, less knowledgeable individuals tend to exhibit more HB (Hertwig, Fanselow, & Hoffrage, 2003); however, even experts are prone to HB when their knowledge is challenged (Pohl, 1992; see Roese & Vohs, 2012, for a review).

An often-overlooked aspect of the uncertainty involved in HB is that the recall of one's initial confidence may be distorted in a similar manner to that of one's answer (see Fischhoff, 1977). For instance, when a crime investigator was initially very confident about the culprit's identity, and then learns she was incorrect, she might recall her initial confidence to be lower than it actually was. Similar errors arise in medical diagnosis (e.g., Littlefair et al., 2016). Doctors have no other choice but to base

decisions on their confidence that there is enough information to provide a reliable diagnosis. Metacognitive hindsight bias (MC-HB) would be a case in which a doctor's confidence is dominated by the outcome of the most recent examination rather than by integrating the overall accumulated data.

MC-HB reflects a shifted recall of confidence based on information provided after the initial answering. Such biased metacognitive monitoring processes are expected to mislead metacognitive control—decisions about collecting additional information or how to use the information already collected (e.g., Metcalfe & Finn, 2008, in a learning context). HB research is rarely informed by metacognitive analysis of factors that might affect people's decisions and the provided answers. In the current research, we explore MC-HB and elucidate conditions particularly prone to MC-HB.

To that end, participants answered difficult knowledge questions and rated the confidence in their answers. Later, after learning the correct answers to half the items, participants tried to recall their original answers as well as their original confidence ratings. MC-HB would occur if the recalled confidence ratings shift depending on learning the correct answers. That is, recalled confidence would retrospectively increase if one learned that one's original answer was correct; recalled confidence would decrease if one learned that one's original answer was incorrect.

Measuring Metacognitive Hindsight Bias

Some factors affect confidence; For instance, familiarity of the question increases confidence (Werth & Strack, 2003). However, such variations do not imply how the correspondence between confidence and actual accuracy of the recalled answers is affected by learning the correct answer. The metacognitive literature often considers two main measures of confidence accuracy, *calibration* and *resolution*. Calibration is the gap between mean confidence and overall success rate (percentage of correct answers out of all answers), measured as over- or under-confidence. Resolution is the discrimination between incorrect and correct answers. Several measures of resolution have been considered in the literature throughout the years (e.g., Fleming & Lau, 2014; Hourihan, Fraundorf, & Benjamin, 2017; Mengelkamp

& Bannert, 2010). Despite methodological critiques (see Fleming & Lau, 2014; Masson & Rotello, 2009), Gamma correlation is the most common measure of resolution, in particular in the case of open-ended tasks (Nelson, 1984; e.g., Tullis, 2018; Yan, Bjork, & Bjork, 2016). In addition to the traditional Gamma correlation, we examined also its recent variation that aims to solve some of its limitations (Higham & Higham, 2019). Perfect resolution (1.0) results from lower confidence in all incorrect answers than in all correct answers. Support for MC-HB arises when recalled confidence is more consistent with the accuracy of the recalled answers than the initial confidence was associated with the initial answers, measured by either calibration or resolution. Notably, a general increase in confidence (e.g., Hertwig, Gigerenzer, & Hoffrage, 1997; Werth & Strack, 2003) may affect calibration but is not directly associated with resolution. This is because if the distribution of confidence ratings is limited to a partial range of values (e.g., between 60% and 80%), the entire distribution can shift upwards by 10% without affecting resolution. This is also the case when people keep their relative confidence, but reduce variability (e.g., confidence increases from 60-80 to 75-85). Only when confidence reaches floor or ceiling levels (e.g., having many answers, both incorrect and correct, provided with 100%), the shift in confidence reduces resolution.

Some studies have measured confidence using Likert scales with verbal titles, rather than percentage scales (e.g., Hom & Ciaramitaro, 2001). One cannot examine calibration with this procedure, because the different units used for confidence and success rates prevent calculating the gap between the two. For resolution, one may use both Likert scales and percentage scales.

Indications for Metacognitive Hindsight Bias (MC-HB)

Several findings in the literature provide indications for MC-HB. Fischhoff (1977) used two-alternative forced choice trivia questions with a percentage scale (0-100%) to measure the subjective probability of an answer being correct. He used both hypothetical and memory-based probability assessment and found HB for both. In particular, the memory group assessed probability of an experimenter-provided answer option being correct and then learned the correct answers to some of the

questions. When participants later did not remember exactly the probabilities they had assigned prior, knowledge of the correct answer led them to overestimate the probabilities they initially assigned to the experimenter-provided answers. This was one of the first demonstrations of HB involving a type of confidence judgment. Notably, though, MC-HB assessment with both calibration and resolution could not be done in that study, because the participants did not generate an answer based on their own knowledge. Rather, they assessed the probability that an experimenter-provided answer was correct. This procedure precludes experience-based heuristic cues which are known to inform metacognitive judgments referring to one's own answers (e.g., based on the gut feeling of fluency, Kelley & Jacoby, 1996).

Winman, Juslin, and Björkman (1998) examined calibration in foresight and hindsight by two-alternative forced choice questions from several domains (e.g., weight comparisons) and collected confidence ratings in percentages. In the foresight phase, participants chose one of two options (e.g., which weight is heavier) and rated their confidence (50-100%). In the hindsight phase, one week later, participants learned the correct answers and hypothetically indicated (1) which option they would have chosen had they not known the correct answer; and (2) the confidence they would have attached to their answer (50%-100%). The authors found that people who were overconfident in foresight became better calibrated in hindsight. This is a demonstration of MC-HB for calibration, but without explicit instruction to recall one's own initial confidence.

Hoch and Loewenstein (1989) focused on resolution in hindsight. They used general-knowledge questions in both two-alternative forced choice and open-ended formats. Like Winman et al. (1998), they examined HB using hypothetical answering after participants learned the correct answers. Hoch and Loewenstein found MC-HB for resolution—better resolution in hindsight than when participants answered the questions and rated confidence based on their own naïve knowledge.

Winman et al. (1998) and Hoch and Loewenstein (1989) showed MC-HB in a hypothetical answering design. Notably, HB in hypothetical and recall designs may involve different underlying processes (Higham, Neil, & Bernstein, 2017).

Therefore, in Experiment 1, we examined whether MC-HB generalizes from a hypothetical HB task to recalling one's own answers with confidence ratings. We also examined MC-HB for both calibration and resolution, something that no prior study has done. Based on previous research, we expected memory for the answers and confidence to be biased, but confidence accuracy—both calibration and resolution—to improve in hindsight. If supported, this would indicate MC-HB for both calibration and resolution in the same study.

Social Effects on Metacognitive Processes

Answering questions in real-life scenarios which involve recall (e.g., in daily conversation or forensic investigation) often carries social considerations of communication norms which are not relevant in hypothetical answering (Grice, 1975). Indeed, one explanation for HB relates to social desirability: Participants may try to appear smarter by giving responses in hindsight which are more consistent with the solution or outcome than their initial knowledge could support (Campbell & Tesser, 1983). However, studies tend to favor cognitive over social factors, in particular memory updating over social desirability, as the basis for HB (see Pezzo, 2011, for a review). Nonetheless, social desirability may still affect MC-HB.

Researchers rarely explore social considerations when examining factors affecting metacognitive monitoring. Exceptions include Karabenick (1996), who found that questions raised by co-learners affected participants' judgment of comprehension. Eskenazi et al. (2016) found that presenting a face randomly gazing towards or away from the answer chosen by the participants affected confidence in their answers, while a car directed similarly to one of the answer options did not affect confidence (see also Jacquot et al., 2015).

Some metacognitive studies have used numerical answers provided by intervals (e.g., "When did the US leave Saigon after the Vietnam war?" Answer: 1970-1980), while framing the task as one of answering interested others. This framing highlights social considerations. In particular, answering in social contexts encourages presenting answers that are useful and informative rather than correct but

ridiculously coarse (e.g., sometime in the last century; Goldsmith, Koriat, & Weinberg-Eliezer, 2002). Recent studies have framed the task as answering a friend to increase undergraduate student involvement (see Ackerman & Goldsmith, 2008). Sidi, Ackerman, and Erez (2018) adapted the procedure used by Ackerman and Goldsmith (2008) for examining whether positive affect influenced confidence calibration. They found better answers but larger overconfidence under positive affect than under neutral affect. Relevant to the present study is their finding that the increased overconfidence under positive affect disappeared when the person who presented the question was a concrete person rather than an imagined friend. These findings hint that when people monitor their own cognitive performance, they include social considerations.

In the present work, we tested the hypothesized MC-HB—distorted recall of confidence ratings in hindsight leading to better calibration (reduced overconfidence) and more reliable resolution in retrospect than participants were in their initial answers. We used the memory design for HB described above to study MC-HB when answering knowledge questions with a social framing adapted from Ackerman and Goldsmith (2008). In Experiment 1 we examined MC-HB regarding one's own answers when answering a friend. In Experiment 2, we delved further into effects of social scenarios on MC-HB.

Experiment 1

In Experiment 1, participants answered the same general-knowledge questions three times in a procedure including four phases. Figure 1 illustrates the procedure with two particular questions, one about Woody Allen and another about Barack Obama, and corresponding confidence ratings.

Phase 1: Initial answer. In this phase, participants answered all questions based on their general knowledge and provided their initial confidence. Let us assume that Liz answered Allen's age when he married for the first time as 20-30 with 60% confidence. Let us also assume that Liz answered the year of Obama's first election as 2000-2005 with 80% confidence. We considered an answer as correct if the actual

answer fell within the provided interval, including both end points. Thus, both of Liz's answers were incorrect.

Phase 2: Learn. Following Ackerman and Goldsmith (2008), we used an incidental learning task for half the questions. Figure 1 demonstrates this phase with the Obama question. This participant did not learn the correct answer to the Allen question, but learned the correct answer to the Obama question which was 2009. By this procedure, we generated a knowledge difference between two sets of questions: the Allen question in the non-learned set and the Obama question in the learned set of questions for this participant. We counterbalanced the presented answers by switching the learned question set for half the participants (in this example, the other half learned about Allen and did not learn about Obama).

In Phase 2 participants rated how new a sentence including the correct answer was for them, without explicit instruction to learn the presented information. This phase had two functions: (a) focus participants' attention on the provided information and (b) guide them to discriminate between their initial answer and what they learned in this phase. To the best of our knowledge, this incidental learning procedure is novel for HB research. Participants' responses in Phase 2 were not our focus.

Phase 3: Recall. This is the *critical phase* for examining HB and MC-HB. It involved recalling the initial answers and the confidence that was associated with each answer in Phase 1. The instructions to Phase 3 included an emphasis on recalling answers and the confidence rating provided in Phase 1. We calculated accuracy of the recalled answer in the same way as the accuracy of the initial answer. The main comparison for examining HB was between one's initial answers in Phase 1 and recall of these answers in Phase 3. The main comparison for examining MC-HB was between one's initial confidence in Phase 1 and recall of one's confidence in Phase 3.

Our 1st hypothesis was that people show HB when recalling their initial answers, replicating numerous studies. Our HB hypothesis was that people would recall their initial answers reliably without learning new information (like the Allen question in Figure 1) and shift toward the correct answer only for those questions to

which they learned the answers (like the Obama question in Figure 1). The latter shift should occur especially when initial answers were incorrect. In our example above, our 1st hypothesis predicted that Liz would show HB by correctly recalling her initial answer to the unlearned Allen question, 20-30, while shifting her recalled answer to the learned Obama question toward the correct answer from 2000-2005 to 2003-2008.

Our 2nd and *central hypothesis* regarded MC-HB. We expected people to adjust their confidence to the learned information. In the above example, when Liz learned the correct answer and her initial answer was incorrect, we expected her "recalled" confidence to be lower than it actually was in Phase 1. Recall of confidence in answers to questions that were not included in Phase 2 should not change as much as it should for questions that were included in Phase 2. This is not trivial, because metacognitive researchers have shown that familiarity of question terms raises metacognitive judgments in various contexts (e.g., Foster, Huthwaite, Yesberg, Garry, & Loftus, 2012; Reder & Ritter, 1992). If our central hypothesis is correct, it means that people can recall (or reassess and derive the same result as) their initial confidence; however, learning the correct answer systematically distorts people's recall of their initial confidence upwards for correct initial answers and downwards for incorrect answers. This is the hypothesized manifestation of MC-HB. In our example above, our 2nd hypothesis predicted that Liz showed MC-HB by recalling her initial confidence in answer to the unlearned question reliably as 60%, while lowering her recalled confidence in answer to the learned question from 80% to 70%. The latter reflects the fact that the current answer is still incorrect, although closer to the correct answer than the initial answer was. We used several measures to examine MC-HB, including mean confidence, calibration, resolution, and confidence in incorrect and correct answers separately. We predicted all these measures to be affected by shifts in recalled confidence ratings. In particular, across answers included in Phase 2 we expected improved recalled answers' accuracy, calibration (reduced overconfidence), and resolution. Resolution would improve by (1) a downward shift in confidence in initially (Phase 1) incorrect answers, as was the case with Liz's recalled answer to the Obama question, and (2) an upward shift in confidence in initially correct answers. We expected all these changes to occur both

relative to answers that were not included in Phase 2 and relative to the initial answers provided by participants in Phase 1. Thus, our main analyses involved analysis of variance (ANOVA) examining the interaction of phase (1 vs. 3) and learning (non-learned vs. learned answers) in Phase 2 on performance, calibration, and resolution when recalling the initial answers in Phase 3.

Phase 4: Re-answer. This phase is a manipulation check for learning in Phase 2 and thus secondary to our main goals. We included it to examine how well participants learned information provided in Phase 2. Here participants re-answered the questions based on both their initial knowledge and the information they learned in Phase 2. We measured accuracy as in Phase 1 and Phase 3, by inclusion of the correct answer in the provided interval. We hypothesized that Liz would answer in Phase 4 again that Allen was 20-30 when he got married, because her knowledge was not changed in the course of the entire answering procedure. Conversely, she would answer that Obama was elected in 2005-2010, reflecting the fact that she had incorporated the knowledge acquired in Phase 2.

Our stimuli included difficult questions to allow room for uncertainty (confidence variability) and meaningful learning. We used fixed intervals (e.g., "provide a 5-year interval"), rather than open intervals ("from ____ to ____"). This provided a constant objective chance of success across experimental phases, to which the confidence ratings refer. Allowing free intervals would affect confidence in ways which are outside our research scope.

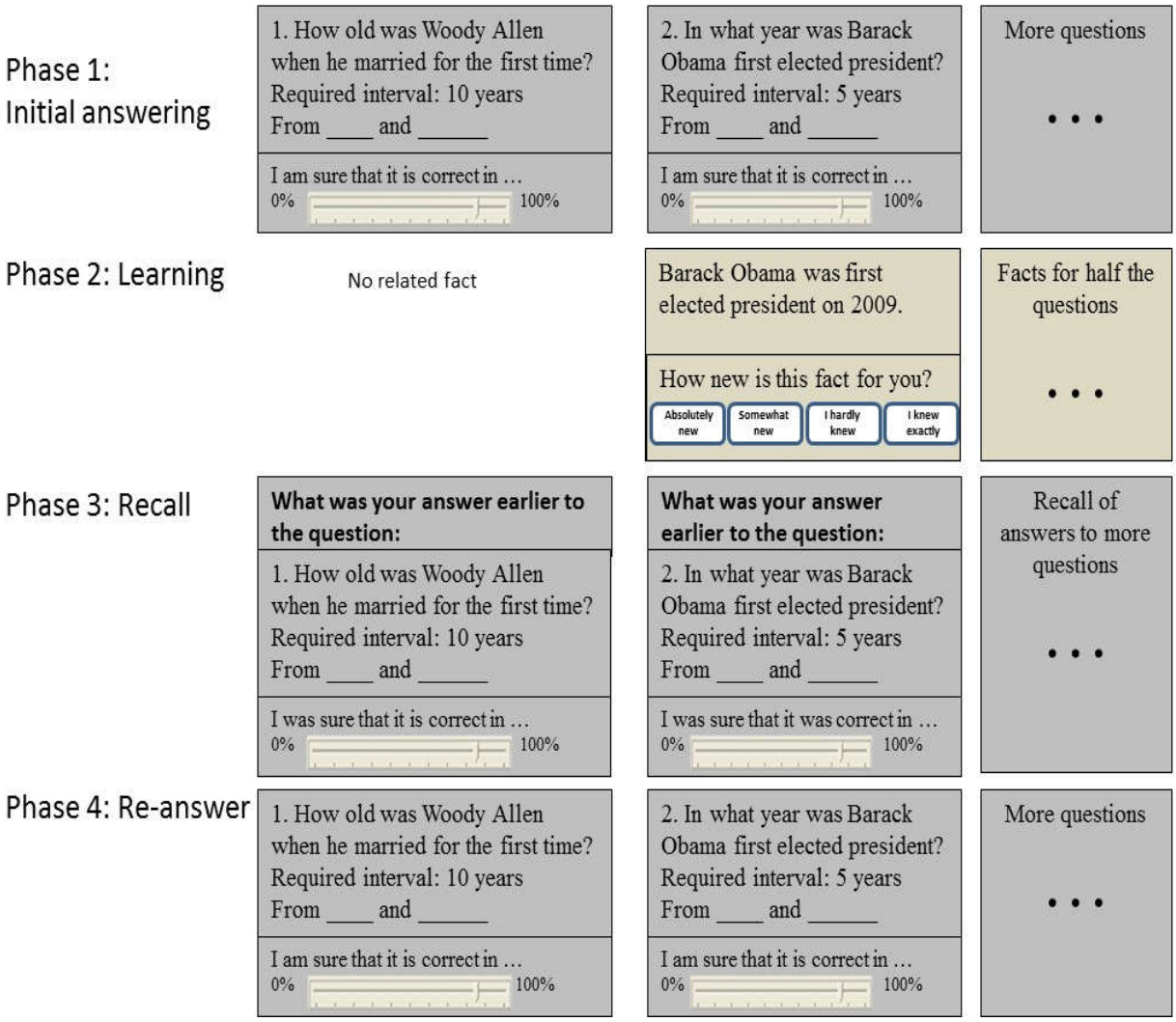

Figure 1. Illustration of the experimental procedure in Experiment 1 for the confidence group. The control group had the same procedure without the confidence ratings.

Method

Participants. Thirty-one undergraduate Technion students (27 received credit; the rest received 10$ for participation; 35% females). We determined sample size to be somewhat larger than previous studies that used a similar paradigm (N = 24-27 per group, Ackerman & Goldsmith, 2008; Sidi et al., 2018) for making sure enough participants pass a selection criterion of variability in accuracy and confidence ratings (not all answers are incorrect or all get the same confidence rating, e.g., 70%). The a-priori required sample size by G*Power for our central comparisons in

Experiment 1, between Phase 1 and Phase 3 within participants (matched pairs), is $N = 27$ for power of 80% and $N = 36$ for power of 90%.

Materials. We determined each question's difficulty and fixed interval via pilot testing, with a different sample ($N = 30$) from the same population, following Ackerman and Goldsmith's (2008) procedure. We chose 32 general-knowledge questions (plus 2 practice questions) with fixed-interval responses that allowed low success rates (15% - 55%) without generating an illusion of success (overconfidence around 15%). The intervals were set for each question by Sidi et al. (2018) based on pretesting.

Design. The experiment followed a 2 (Phase: initial answer, recall) × 2 (Learning: non-learned, learned) within-subjects design.

Procedure. The experiment occurred in a small lab with 2-8 participants in each session. Participants learned that the experiment included four phases, but participants received the relevant instructions just before each phase. Participants answered the questions as if an interested friend had asked and wanted sincere answers. Questions appeared in a new random order in each phase.[1]

Results

We excluded two participants who had no correct answers or had no variability in confidence ratings in at least one phase. The results, therefore, are based on 29

[1] Ackerman and Goldsmith (2008) compared two groups which answered with and without confidence ratings and found no difference between these tasks. However, in other contexts, some studies have found performance differences between answering with and without confidence ratings (e.g., Double & Birney, 2017; Petrusic & Baranski, 2003; Soderstrom, Clark, Halamish, & Bjork, 2015). In all phases of Experiment 1, we compared one group which answered with confidence ratings to another comparable group ($N = 29$) which performed the entire task without confidence ratings, for determining whether eliciting confidence affects performance in the task we used. Replicating Ackerman and Goldsmith's (2008) findings, eliciting confidence did not affect answering in any phase. This finding suggests that people provide and recall the answer and their confidence as two separate units of information. We used data from the confidence group to examine the hypothesized MC-HB.

participants. Descriptive results and simple effects appear in Table 1. Notably, there were no floor or ceiling effects: Performance and confidence were significantly above zero and below 100%, *p*s < .0001, providing adequate variability for all measures in all phases.

Manipulation check for the learning phase. In Phase 4, the re-answer phase, participants more accurately answered questions for which they learned the correct answers in Phase 2 (learn phase) than they answered them in Phase 1 (initial-answering phase). There were few changes between the phases for non-learned answers in performance and metacognitive measures. See Table 1.

Hindsight bias (HB). To test our 1st hypothesis, we examined HB by analyzing success rates in the initial answer phase and recall phase (see Figure 2). A within-subject ANOVA of Phase (initial answer vs. recall) × Learning (non-learned vs. learned) revealed a marginal effect of phase, $F(1, 28) = 3.62$, MSE = 157.11, $p = .067$, $\eta_p^2 = .114$, no effect of learning, $F(1, 28) = 2.57$, MSE = 686.43, $p = .120$, $\eta_p^2 = .084$, but a phase-by-learning interaction, $F(1, 28) = 4.63$, MSE = 191.82, $p = .040$, $\eta_p^2 = .142$. While success rates for the non-learned answers did not change across the two phases, $t < 1$, success rates for learned answers improved significantly when participants recalled their initial answers, $t(28) = 2.41$, $p = .023$, Cohen's $d = 0.45$. See dashed lines in Figure 2. See reports of simple effects comparing non-learned and learned items in Table 1. These findings reflect the classic HB.

Table 1. Means (SDs) of answers by the confidence group of Experiment 1 ($N = 29$) in the three answering phases, with significance of differences between non-learned and learned answers and among the three answering phases.

Measure	Non-learned	Learned	Paired t value
Phase 1 – Initial Answer			
Success rate (%)	28.7 (14.5)[a]	31.0 (14.2)[a]	< 1
Confidence (%)	42.6 (17.8)[a]	44.2 (17.7)[a]	1.3
Calibration (overconfidence)	13.9 (17.0)[a]	13.2 (18.3)[a]	< 1
Resolution	.29 (.42)[a]	.28 (.40)[a]	< 1
Phase 3 – Recall			
Success rate (%)	28.5 (13.7)[a]	35.9 (17.8)[b]	2.3 *
Confidence (%)	45.7 (19.5)[b]	51.0 (20.5)[b]	2.6 *
Calibration (overconfidence)	17.2 (20.6)[a]	15.1 (18.9)[a]	< 1
Resolution	.31 (.37)[a]	.56 (.30)[b]	3.7 **
Phase 4 – Re-answer			
Success rate (%)	30.1 (14.7)[a]	76.3 (17.3)[c]	12.6***
Confidence (%)	48.0 (19.3)[b]	82.9 (14.0)[c]	12.1 ***
Calibration (overconfidence)	17.9 (21.2)[a]	6.6 (11.1)[b]	3.1**
Resolution	.39 (.34)[a]	.68 (.43)[b]	3.0 **

* $p < .05$. ** $p \leq .01$, *** $p \leq .0001$ for differences between non-learned and learned answers

Note. Different superscripts within the same column denote significant differences between phases ($p < .05$).

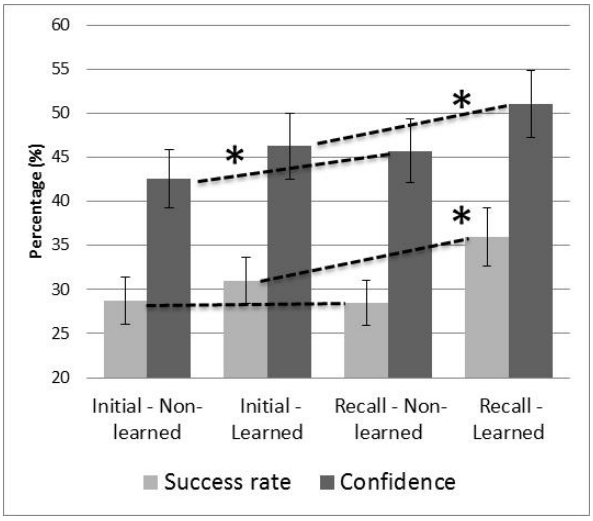

Figure 2. Experiment 1 - Success rates, confidence, and calibration (overconfidence) for the initial answer phase (Phase 1) and for the recall phase (Phase 3), for non-learned and learned answers. The difference between adjacent success rate and confidence bars represents overconfidence. Error bars represent standard error of the means. The dashed lines at the top of the success rate bars represent the interaction indicative of Hindsight Bias. * represents a significant difference between the ends of the closest dashed line, $p < .05$.

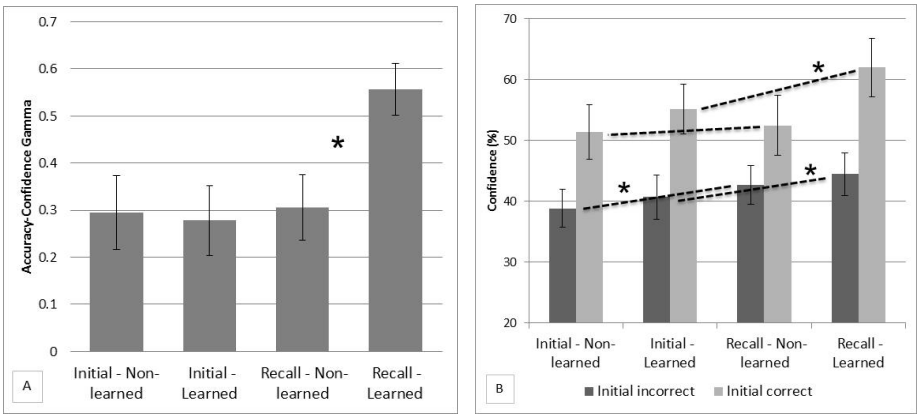

Figure 3. Experiment 1 - A – Resolution for the initial answer phase (Phase 1) and for the recall phase (Phase 3), for non-learned and learned answers. B – Confidence in answers which were incorrect and correct in Phase 1. The dashed lines emphasize the source of the three-way interaction. Error bars represent standard error of the means. * represents a significant difference, $p < .05$.

Metacognitive hindsight bias (MC-HB)

Our 2nd and main hypotheses regarded confidence, calibration, and resolution. See means in Table 1.

Confidence. ANOVA as above, of Phase (initial answer vs. recall) × Learning (non-learned vs. learned), on confidence revealed two main effects. Participants were more confident in the recall phase than in the initial-answering phase, $F(1, 28) = 8.17$, MSE = 55.02, $p = .008$, $\eta_p^2 = .226$, and were more confident in learned answers than in non-learned answers, $F(1, 28) = 6.91$, MSE = 84.27, $p = .014$, $\eta_p^2 = .198$. No interaction emerged, $F = 1.2$, MSE = 21,20, $p = .278$, $\eta_p^2 = .042$. We expected confidence in incorrect answers to decrease. This prediction was not supported. The general increase in confidence can be explained by familiarity of the questions themselves (see General Discussion).

Calibration. Participants were overconfident in all phases, all *p*s < .0001 (see Figure 2). ANOVA on overconfidence, as above, yielded no significant effects, $F < 1$ for both main effects and $F = 2.04$, MSE = 85.48, $p = .164$, $\eta_p^2 = .068$, for the interaction. Thus, counter to our prediction, there was no MC-HB for calibration (see Figure 2): Learning did not significantly affect the extent of overconfidence.

Resolution. Resolution was reliable for all phases, all *p*s ≤ .001, indicating robust discrimination between incorrect and correct answers (see Figure 3A and Table 1). Most importantly, examination of resolution revealed MC-HB in our memory design, replicating previous results obtained with hypothetical answering (Hoch & Loewenstein, 1989). ANOVA as above on resolution revealed two marginal main effects, $F(1, 28) = 4.20$, MSE = 0.144, $p = .05$, $\eta_p^2 = .130$ for phase, $F(1, 28) = 3.39$, MSE = 0.117, $p = .076$, $\eta_p^2 = .108$ for learning, and an interaction, $F(1, 28) = 9.47$, MSE = 0.055, $p = .005$, $\eta_p^2 = .253$. Resolution did not change for non-learned answers, $t < 1$. Conversely, resolution improved for learned answers,

$t(28) = 3.67$, $p = .001$, $d = 0.682$. Thus, learning improved discrimination between incorrect and correct answers in hindsight.[2]

To further examine this discrimination, we examined confidence in incorrect and correct answers. A 3-way within-subject ANOVA of Phase (1st vs. 3rd) × Learning (non-learned vs. learned) × Initial answer accuracy (incorrect vs. correct) on confidence revealed three main effects and a three-way interaction. The first two main effects were reported above. Additionally, there was a main effect of initial answer accuracy, $F(1, 28) = 29.23$, MSE = 10678.04, $p < .0001$, $\eta_p^2 = .511$, which reflects overall good discrimination between incorrect and correct answers. The three-way interaction, $F(1, 28) = 5.19$, MSE = 125.35, $p = .030$, $\eta_p^2 = .156$, stemmed from increased recalled confidence in initially incorrect answers, $F(1, 28) = 5.88$, MSE = 417.70, $p = .022$, $\eta_p^2 = .174$, regardless of learning (dashed lines in Figure 3B). For initially correct answers, participants succeeded in recalling their confidence ratings in non-learned correct answers, reporting equivalent mean confidence to its level in Phase 1, $t < 1$; however, participants became more confident in initially correct answers that they learned, $t(28) = 3.71$, $p = .001$, $d = 0.69$, explaining the increased discrimination between incorrect and correct learned answers (see Figure 3B).

The above resolution analyses were based on our assumption that people recall their answer and their confidence as a single information unit. Thus our resolution measure referred each confidence rating to the answer with which it was provided. However, we directed participants to recall their confidence in the previous answering block. If participants recalled their answers and their confidence ratings as two separated information units, it is possible that the recalled confidence is in fact associated with the initial answer rather than with the recalled one. Thus, we calculated resolution also as the gamma correlation between the recalled confidence

[2] Recently, Higham and Higham (2019) suggested a method to overcome some of the problems with Gamma correlation by using the area under the ROC curve (AUC) associating confidence and accuracy for each participant. Their improved Gamma = (2*AUC)-1. Using this measure yielded MC-HB for resolution as well. ANOVA as above revealed a marginal effect for the phase, $F(1, 28) = 3.37$, MSE = 0.09, $p = .077$, $\eta_p^2 = .107$, a significant main effect for learning, $F(1, 28) = 4.43$, MSE = 0.12, $p = .044$, $\eta_p^2 = .137$, and a significant interaction, $F(1, 28) = 11.39$, MSE = 0.11, $p = .002$, $\eta_p^2 = .289$.

and the correctness of the initial answers. These correlations were both positive, (*M*non-learned = .26, *SD* = .36, $t(28) = 3.98$, $p < .0001$; *M*learned = .43, *SD* = .48, $t(28) = 4.90$, $p < .0001$). ANOVA examining the interaction of resolution types (recalled confidence with recalled vs. initial answers) and Learning (non-learned vs. learned) revealed a main effect of resolution types, $F(1, 28) = 4.41$, $MSE = 0.20$, $p = .045$, $\eta_p^2 = .136$, indicating that resolution was significantly stronger when associating confidence with the recalled answers (*M*non-learned = .31, *SD* = .37; *M*learned = .56, *SD* = .30) than with the initial answers. A main effect of feedback emerged as well, $F(1, 28) = 9.80$, $MSE = 1.29$, $p = .004$, $\eta_p^2 = .259$. Most importantly, there was no interactive effect, $F(1, 28) = 1.40$, $MSE = 0.05$, $p = .247$, $\eta_p^2 = .047$. This finding indicates that MC-HB is consistent across the two resolution types. The stronger resolution when associating the recalled confidence with the recalled answers supports our assumption that participants did not ignore their recalled (and biased by HB) answers when reporting the recalled confidence.

Overall, Experiment 1 generalizes the classic HB to recalling one's own answers in a social scenario. Additionally, it demonstrates pronounced MC-HB in resolution, but not in calibration. Experiment 2 examines the robustness of these findings and delves further into the social aspects of MC-HB.

Experiment 2

To understand better the role of social considerations in MC-HB, we manipulated the social scenario in which answering and recalling occurred. Our 3rd hypothesis was that the more emphasis given to other people considering the value of one's answers and confidence, the larger is the MC-HB when one tries to recall one's initial answers in Phase 3. The rationale was that people use recalled confidence in hindsight to justify their initial answers. People also use recalled confidence in hindsight to avoid losing face for being confident in incorrect answers or being uncertain about correct answers. That is, we expected the polarization in confidence ratings in hindsight described above to be stronger when participants expected someone else to review their initial answers than when participants

provided their initial answers anonymously. We expected this increased polarization to appear as improved resolution in social contexts relative to anonymous situations.

There were three groups in Experiment 2. The *Anonymous* group answered after we assured them that nobody would be able to link them to their answers (see Deutsch & Gerard, 1955, for a similar procedure). We asked the *Imagined-Friend* group to answer an imagined friend, replicating Experiment 1's procedure. Finally, we strongly emphasized the social scenario in the *Peer-Review* group. We told participants in this group that in the final phase of the experiment they would review the answers of another participant sitting with them in the room, and that one of their peers would review their own answers.

Method

Participants. Ninety-three Technion undergraduates (88% for credit; 35% females) were randomly assigned to the three groups. In this experiment we had a mixed design, with three groups and a within-participant comparison between Phase 1 and Phase 3 (two measurements). The a-priori sample size calculated by G*Power is N = 42 for power of 80% and N = 54 for power of 90%.

Materials, design and procedure.

The materials, design, and procedure were identical to Experiment 1, including initial answer, learn, recall, and re-answer phases. The differences were in the instructions provided at the beginning of the experiment and an extra phase for the two social groups, as described below. The specific part of the initial instructions for the anonymous group included: "The research results will be analyzed for all participants together, in a way that even the researchers will not be able to identify you." The specific part of the instructions for the imagined-friend group was: "When you answer a question, imagine that a friend who does not know the answer asked you this question, because he would like to know and you try to help him. The study includes several phases. At the final phase, you will see answers of a participant who took part in this study in the past. In that phase, you will be asked to assess how helpful that participant's answers were." The peer-review group received the same

initial instructions, except for this sentence: "At the final phase, you will review the answers of one of the participants sitting in the room with you (and one of them will review your answers)." At this additional 5th and final phase, the *review phase*, participants of both social groups in fact reviewed 10 answers of the same past participant. All participants rated confidence in all phases, except for the review phase. Experiment 2 followed a 2 (Phase: initial answer, recall) × 2 (Learning: non-learned, learned) x 3 (Group: anonymous, imagined friend, peer review) mixed design with Group as the between-subjects factor.

Results

Using the criteria from Experiment 1, we excluded three participants, one participant from each group. The manipulation checks for learning effectiveness replicated all the related findings of Experiment 1.

Hindsight bias (HB). We repeated the ANOVAs of Phase (initial answer vs. recall) × Learning (non-learned vs. learned) on performance from Experiment 1 with Group (anonymous vs. imagined friend vs. peer review) as an additional factor. All results replicated (see Figure 4). The main effect of phase was significant, $F(1, 87) = 12.02$, MSE = 771.19, $p = .001$, $\eta_p^2 = .121$, the main effect of learning was not significant, $F(1, 87) = 2.14$, MSE = 506.81, $p = .147$, $\eta_p^2 = .024$, and there was an interaction between the two, $F(1, 87) = 19.65$, MSE = 991.68, $p < .0001$, $\eta_p^2 = .184$. For the non-learned answers, there was no significant difference between the phases, $t < 1$, while for the learned answers there was an improvement in Phase 3 relative to Phase 1, $t(89) = 4.29$, $p < .0001$, $d = 0.45$. Importantly, there were no interactions with group, all Fs < 1.2. Thus, we again observed robust HB, but the social scenario did not affect HB.

Metacognitive hindsight bias (MC-HB)

Confidence. A mixed three-way (Phase, Learning, Group) ANOVA on confidence revealed somewhat different results from Experiment 1 (see Figure 4). The main effect of phase was now not significant, $F(1, 87) = 2.79$, MSE = 164.44, $p = .10$, $\eta_p^2 = .031$, although in the same direction, while the main effect of learning

remained significant, $F(1, 87) = 4.87$, MSE = 371.182, $p = .030$, $\eta_p^2 = .053$. More importantly, our larger sample revealed a strong phase-by-learning interaction, $F(1, 87) = 14.26$, MSE = 303.92, $p < .0001$, $\eta_p^2 = .141$. Here, confidence did not change for non-learned answers, $t < 1$, but increased for learned answers, $t(89) = 2.91$, $p = .005$, d = 0.31. Finally, there were no group effects, all ps $\geq .10$.

Calibration. The increase in confidence after learning was smaller than the increase in success rates. This resulted in improved calibration (reduced overconfidence). ANOVA as above on overconfidence revealed no main effects, all Fs < 1. However, the ANOVA did reveal a phase-by-learning interaction, $F(1, 87) = 4.40$, MSE = 197.62, $p = .039$, $\eta_p^2 = .048$. Overconfidence did not change for non-learned answers, $t < 1$, but was smaller for learned answers, $t(89) = 2.28$, $p = .025$, $d = 0.24$. This was the case regardless of group, all Fs < 1. Thus, we found MC-HB for calibration, which showed only a trend in Experiment 1. This MC-HB for calibration occurred regardless of the social scenario, and generalizes previous findings derived from the hypothetical answering paradigm (Winman et al., 1998).

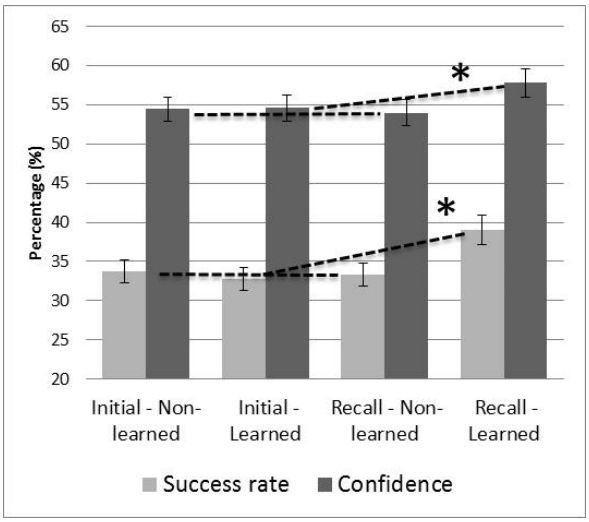

Figure 4. Experiment 2 - Success rates, confidence, and calibration (overconfidence) for the initial answer phase (Phase 1) and for the recall phase (Phase 3), for non-learned and learned answers, across the three groups. The difference between adjacent success rate and confidence bars represents overconfidence. Error bars represent standard error of the means. The dashed lines represent the interaction indicative of Hindsight Bias, in success rates, and Metacognitive Hindsight Bias, in confidence. * represents a significant difference, $p < .05$.

Resolution. Resolution results of Experiment 2 reveal more decisive main effects than in Experiment 1 and replicated the interaction between phase and learning (Figure 4). ANOVA yielded a significant main effect of phase, $F(1, 87) = 4.06$, MSE = 0.16, $p = .047$, $\eta_p^2 = .045$, no effect of learning, $F < 1$, and a phase-by-learning interaction, $F(1, 87) = 8.23$, MSE = 0.42, $p = .005$, $\eta_p^2 = .086$. For the non-learned answers, there was no significant difference between the phases, $t < 1$, while for the learned answers there was an improvement in Phase 3 relative to Phase 1, $t(89) = 3.18$, $p = .002$, $d = 0.34$.

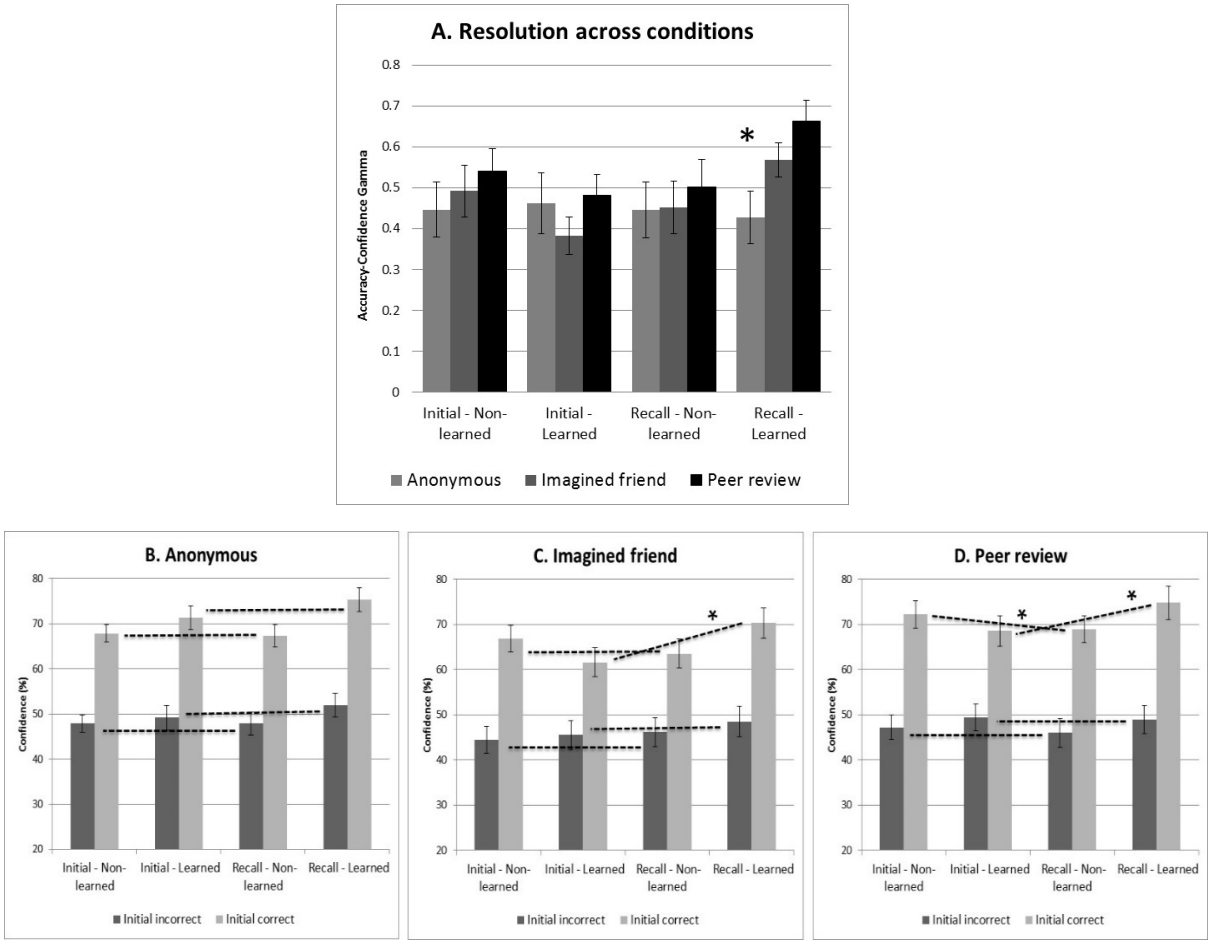

Figure 5. Experiment 2 – A. Resolution for the initial answer phase (Phase 1) and for the recall phase (Phase 3), for non-learned and learned answers. B, C, and D – Confidence in answers which were incorrect and correct in Phase 1 in each condition. The dashed lines emphasize the source of the interaction. Error bars represent standard error of the means. * represents a significant difference, $p < .05$.

Additionally, there was a three-way interaction with group, $F(2, 87) = 3.25$, MSE = 0.17, $p = .044$, $\eta_p^2 = .069$ (see Figure 5A). To understand the source of this interaction, we conducted a two-way ANOVA as above for each group. The anonymous group showed no effects, all Fs < 1. The imagined-friend group showed a marginal effect of phase, $F(1, 29) = 4.14$, MSE = 0.16, $p = .051$, $\eta_p^2 = .125$, no effect of learning, $F < 1$, and a phase-by-learning interaction, $F(1, 29) = 7.89$, MSE = 0.38, $p = .009$, $\eta_p^2 = .214$. Replicating Experiment 1's results, resolution did not change when recalling confidence in non-learned answers, while resolution improved when recalling confidence in learned answers, $t(29) = 3.47$, $p = .002$, $d = 0.635$. The same pattern emerged for the peer-review group. There was a main effect of phase, $F(1, 29) = 6.08$, MSE = 0.15, $p = .020$, $\eta_p^2 = .173$; no effect of learning, $F < 1$, and a phase-by-learning interaction, $F(1, 29) = 6.28$, MSE = 0.37, $p = .018$, $\eta_p^2 = .178$. Resolution did not change when recalling confidence in non-learned answers, $t < 1$, but improved for the learned answers, $t(29) = 3.77$, $p = .001$, $d = 0.69$. Thus, resolution did not improve in hindsight when participants expected anonymity. Conversely, resolution did improve in hindsight similarly in the two social scenarios, imagined-friend and peer-review[3].

Examining confidence differentiation between incorrect and correct initial answers, as in Experiment 1, we replicated the pattern of higher confidence only for the answers that were initially answered correctly and afterwards affirmed in the learning phase. Notably, though, this was the case only for the social conditions (both interactive effects had $p \leq .001$ and simple effects for these improvements, $p \leq .005$), but not for the anonymous condition, which had $F < 1$ for all effects. See Figure 5B, 5C, and 5D.

[3] Using Higham and Higham's (2019) version of Gamma correlation similarly yielded MC-HB in the two social conditions only. There was a main effect of phase, $F(1, 87) = 6.01$, MSE = 0.20, $p = .016$, $\eta_p^2 = .065$, no main effect of learning, $F < 1$, phase-by-learning interaction, $F(1, 87) = 16.85$, MSE = 0.73, $p < .0001$, $\eta_p^2 = .162$, and a triple interaction with the condition, $F(2, 87) = 3.58$, MSE = 0.154, $p = .032$, $\eta_p^2 = .076$. For the anonymous condition, the phase-by-learning interaction was not significant, F < 1, while it was significant for the imagined-friend group, $F(1, 29) = 8.33$, MSE = 0.34, $p = .007$, $\eta_p^2 = .223$, and for the peer-review condition, $F(1, 29) = 14.25$, MSE = 0.68, $p = .001$, $\eta_p^2 = .33$.

Like in Experiment 1, associating confidence with accuracy in the initial answering phase for the social conditions revealed a highly similar pattern of MC-HB to associating confidence with recalled answers' accuracy: We observed a main effect of learning, $F(1, 58) = 5.17$, MSE = 0.97, $p = .027$, $\eta_p^2 = .082$. In Experiment 1, the resolution calculation associating recalled confidence to accuracy in the recall phase was better than when relating confidence to the accuracy in the initial answering. In Experiment 2, the comparison between the resolution types, confidence relative to initial or recalled answer, revealed no significant differences, $F(1, 58) = 1.46$, MSE = 0.04, $p = .232$, $\eta_p^2 = .025$ for the main effect of resolution type, and $F < 1$ for the interactive effects between resolution type and learning and between resolution type and condition. Thus, despite the inconsistency in the advantage found in Experiment 1 for one resolution type over the other, the results reveal consistency in the main finding of MC-HB on resolution, regardless of the resolution type.

In sum, Experiment 2 replicated most results of Experiment 1. The main difference was that in Experiment 1, we found MC-HB only in resolution. In Experiment 2, we found MC-HB both in calibration and resolution. Also, in Experiment 2, the kind of social scenario did not affect calibration, but did affect resolution. Thus, participants showed improved resolution in hindsight only when the presented social scenario involved considering other people's view on the provided answers.

General Discussion

In this study we considered the possibility that people demonstrate MC-HB. Namely, we examined whether people adjust recalled confidence ratings, as they adjust recalled answers in HB. Participants answered general-knowledge questions during Phase 1 (initial answer), indicating their confidence in each answer. In Phase 2 (learn), they learned the answers to half these questions. In Phase 3 (recall), they tried to recall their Phase-1 answers and the confidence that they initially attached to each answer. In Phase 4 (re-answer), participants re-answered the questions to the best of their ability. Because we found robust HB in our paradigm across all

conditions consistent with the existing HB literature, we focus our discussion on MC-HB.

In Phase 3, our participants had to recall their Phase-1 (initial) confidence. They showed a remarkable ability to recall their confidence when they did not learn the correct answers (Experiment 1 and Experiment 2), for initially-incorrect (Phase-1) answers despite learning the correct answers in Phase 2 (Experiments 1 and Experiment 2), and when we promised anonymity (Experiment 2). It is not clear, though, whether people indeed recall their initial confidence or infer their confidence anew. Two findings in Experiment 1 support inference of confidence anew. First, there was a general rise in confidence in the recall phase relative to the initial confidence in Experiment 1 (see Figure 3B). As mentioned above, this finding may stem from familiarity of the question itself (Foster et al., 2012; Reder & Ritter, 1992; Werth & Strack, 2003). Second, resolution calculation by associating the recalled confidence with the accuracy of the recalled answer was stronger than when associating the same confidence with the accuracy of the initial answers. However, both findings did not replicate in Experiment 2. Nevertheless, in both experiments the quite successful recall, as detailed above, highlights the recall bias reflected in the upwards shift in confidence in Phase 3 regarding initially correct answers. This shift was the main source of the improved resolution for learned items in hindsight. Our findings indicate that this bias is predictable, robust, and generates the MC-HB we found.

Furthermore, we expected recalled confidence in incorrect answers that were learned in Phase 2 to shift downwards. This prediction was not supported: In both experiments, the improvement in resolution in the social conditions stemmed mainly from higher confidence in answers that were initially correct. It seems that familiarity, or other heuristic cues, override the downwards shift we expected. Future research is needed to determine whether the accurate recall of confidence incorporates two different inferential effects that cancel each other, or stems from a reliable direct recall of confidence per se. Multinomial processing tree models of

hindsight bias would be a good way to address this question (see Erdfelder & Buchner, 1998).

Unlike previous studies of MC-HB, we used two measures of monitoring accuracy: calibration and resolution. We also examined social considerations that lead to the systematic shift in recalled confidence in hindsight. For HB itself, social motivation is likely not the main underlying mechanism (Pezzo, 2011). In the present paradigm, we found that answering when anticipating that another person would use or review one's answers affects metacognitive but not cognitive processes: MC-HB increased as one's identifiability increased from anonymity to being reviewed by a peer present in the room, while the social contexts we used did not affect HB. These findings add to the scarce literature on social aspects of metacognitive processes in general, and HB in particular.

We measured MC-HB regarding confidence accuracy by examining changes in calibration and resolution across experimental phases. Overall, we found MC-HB in resolution in all socially-framed conditions, while we did not find MC-HB in resolution when answering anonymously. Campbell and Tesser (1983) demonstrated that HB magnitude correlated positively with individuals' scores on a measure of social desirability. Perhaps MC-HB is also prone to individual differences, such as social desirability. This is a direction for future research. Another direction for future research is to link both HB and MC-HB with the vast literature on perspective taking and theory of mind (see Birch & Bernstein, 2007; Kuhn, 2000).

The data pattern we observed resembles that in prior studies in which researchers have used manipulations meant to selectively affect three different hindsight components: memory distortion, inevitability, and foreseeability (Blank, Nestler, von Collani, & Fischer, 2008; Nestler, Blank, & Egloff, 2010). In one experiment, retention interval and the availability of a misattribution source (e.g., the room's lighting) produced a double dissociation between memory distortion and foreseeability (Nestler et al., 2010). Such dissociations are typically seen as strong evidence for independence among processes, components, or systems (Glanzer & Cunitz, 1966). Later theoretical work integrated the three hindsight bias components

into a hierarchical framework (Roese & Vohs, 2012). According to this framework, memory distortion and inevitability represent cognitive processes whereas foreseeability represents metacognitive processes. We found differential social effects on HB and MC-HB: HB was consistent across social conditions while MC-HB depended on the social condition. Future research should examine such dissociations between factors affecting HB and MC-HB.

Our findings regarding MC-HB for calibration were less clear than our findings regarding MC-HB for resolution. In Experiment 1, where confidence was higher in Phase 3 than in Phase 1 regardless of learning, MC-HB for calibration showed a trend in the expected direction of reduced overconfidence. In Experiment 2, confidence increased for the learned answers only. In this case, MC-HB for calibration was significant and consistent across the social conditions. Perhaps MC-HB for calibration is weaker than MC-HB for resolution, and thus showed significance only with a larger sample. So far, we discussed HB and MC-HB as biases which should be eliminated. However, one may see these shifts in answers and confidence as functional because they stem from incorporating new knowledge into one's semantic net. This view of HB is not new (Hawkins & Hastie, 1990; Hoffrage, Hertwig, & Gigerenzer, 2000), but deserves further consideration from a metacognitive perspective for better understanding MC-HB processes.

As reviewed prior, most research on MC-HB is 30 years old. There is value in reviving this research domain using recent insights and research questions from the metacognitive literature (see also Bernstein et al., 2016). First, Wallace, Chang, Carroll, and Grace (2009) found that HB is weaker when effortful learning is involved than when the answers are given. Studies of learning and problem solving suggest that engagement allows people to improve monitoring accuracy (e.g., Mitchum & Kelley, 2010; Thiede, Anderson, & Therriault, 2003). Specifically, such improvements arise in computerized environments, which seem to be dominated by shallower default processing than traditional paper-and-pencil environments (Lauterman & Ackerman, 2014; Sidi, Shpigelman, Zalmanov, & Ackerman, 2017). Our incidental learning procedure in Phase 2 (asking participants how new the

information is) involved some engagement, which might not be required when the answers are given without a requirement for action. Future research should consider whether this procedure promotes source monitoring—discriminating which information was known beforehand and which was just learned—as we intended it to do.

A second reason to re-evaluate hindsight bias research in light of new developments in metacognitive research refers to the idea that metacognitive monitoring is based on heuristic cues (Koriat, 1997). Specifically, the judgments provided by people who are naïve differ from those provided by people who see the answers (e.g., Kelley & Jacoby, 1996; Rhodes & Tauber, 2011). Hypothetical answering after learning the correct answers eliminates people's experience of naïveté. Thus, it is possible that MC-HB in hypothetical procedures, as used by Winman et al. (1998) and Hoch and Loewenstein (1989), and in recall procedures, as we used here, differ in the heuristic bases for confidence. These bases may also change across experimental phases.

A third reason to re-evaluate hindsight bias research in light of new developments in metacognitive research refers to a dominant topic in metacognitive research: the allocation of answering time. This aspect raises questions regarding HB such as how much time people invest in answering each question depending on its difficulty, familiarity, motivation, time pressure, etc. (see Ackerman & Thompson, 2017; Kornell & Bjork, 2007, for reviews).

A final reason to re-evaluate hindsight bias research as it pertains to metacognition focuses on the applied consequences of HB. Our results show that learning the correct answers distorts memory (HB) but improves resolution and sometimes calibration (MC-HB). Taking this finding from the laboratory to the real world, we offer the following. Consider a student who studies for an exam by testing herself (Bae, Therriault, & Redifer, 2018; Roediger & Karpicke, 2006). While testing herself, the student rates her confidence in each answer. After learning the correct answers, two things happen: (1) The student thinks that she knew the answers prior, even when she did not (HB); and (2) the student improves in terms of

resolution and maybe also calibration (MC-HB). While continuing to study for the exam and studying for future exams, the student focuses on what should be relearned. Consequently, the student's study habits and test performance improve. Accompanying this improvement is continuing and persistent HB for learned answers: "I knew that answer!" What is important here is that the student learns the correct answers and can demonstrate this learning on the actual exam. HB is a by-product of learning in this example, but the distorted confidence in hindsight supports more attuned study effort towards future exams than the initial answering allowed. Future research should focus on the relation between hindsight bias and learning (Bernstein et al., 2016; Henriksen & Kaplan, 2003).

In sum, our findings could improve decision making in contexts such as forensic investigations, education, and medical diagnosis. Biases stemming from MC-HB may affect people's decision-making competence in a variety of social situations and domains.

Acknowledgments

We thank Hartmut Blank, Megan Giroux, Eric Mah and Yael Sidi for helpful comments on a prior draft. This work was supported by the Israel Science Foundation [grant No. 234/18] and by grants from the Canada Research Chairs Program (950-228407) and the Social Sciences and Humanities Research Council of Canada (435-2015-0721).

The data and materials are available by request from the corresponding author.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting--With and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1224-1245.

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences, 21*(8), 607-617.

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2018). Investigating the testing effect: retrieval as a characteristic of effective study strategies. *Learning and Instruction, 60*, 206-214.

Bernstein, D. M., Aßfalg, A., Kumar, R., & Ackerman, R. (2016). Looking backward and forward on hindsight bias. In J. Dunlosky & S. K. Tauber

(Eds.), *The Oxford Handbook of Metamemory* (pp. 289-304). New York, NY: Oxford University Press.
Birch, S. A., & Bernstein, D. M. (2007). What can children tell us about hindsight bias: A fundamental constraint on perspective–taking? *Social cognition, 25*(1), 98-113.
Blank, H., Nestler, S., von Collani, G., & Fischer, V. (2008). How many hindsight biases are there? *Cognition, 106*(3), 1408-1440.
Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An individual difference analysis. *Journal of Personality, 51*(4), 605-620.
Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology, 51*(3), 629-636.
Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's progressive matrices. *Thinking & Reasoning, 23*(2), 190-206.
Dror, I. E., Morgan, R. M., Rando, C., & Nakhaeizadeh, S. (2017). The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making. *Journal of forensic sciences, 62*(3), 832-833.
Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 387-414.
Eskenazi, T., Montalan, B., Jacquot, A., Proust, J., Grèzes, J., & Conty, L. (2016). Social influence on metacognitive evaluations: The power of nonverbal cues. *The Quarterly Journal of Experimental Psychology, 69*(11), 2233-2247.
Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 288-299.
Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 349-358.
Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience, 8*(443). doi:10.3389/fnhum.2014.00443
Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta Psychologica, 139*(2), 320-326.
Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior, 5*(4), 351-360.
Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General, 131*(1), 73-95.
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics* (pp. 41–58). New York: Academic Press.
Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological bulletin, 107*(3), 311-327.

Henriksen, K., & Kaplan, H. (2003). Hindsight bias, outcome knowledge and adaptive learning. *BMJ Quality & Safety, 12*(suppl 2), ii46-ii50.
Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory, 11*(4-5), 357-377.
Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological review, 104*(1), 194-202.
Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods, 51*(1), 108-125.
Higham, P. A., Neil, G. J., & Bernstein, D. M. (2017). Auditory hindsight bias: Fluency misattribution versus memory reconstruction. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1144-1159.
Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 605–619.
Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 566-581.
Hom, H. L., & Ciaramitaro, M. (2001). GTIDHNIHS: I knew-it-all-along. *Applied Cognitive Psychology, 15*(5), 493-507.
Hourihan, K. L., Fraundorf, S. H., & Benjamin, A. S. (2017). The influences of valence and arousal on judgments of learning and on recall. *Memory & cognition, 45*(1), 121-136.
Jacquot, A., Eskenazi, T., Sales-Wuillemin, E., Montalan, B., Proust, J., Grèzes, J., & Conty, L. (2015). Source unreliability decreases but does not cancel the impact of social information on metacognitive evaluations. *Frontiers in Psychology, 6*, 1385.
Karabenick, S. A. (1996). Social influences on metacognition: Effects of colearner questioning on comprehension monitoring. *Journal of Educational Psychology, 88*(4), 689-703.
Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*(2), 157-175.
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349-370.
Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224.
Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*(5), 178-181.
Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior, 35*, 455-463.

Littlefair, S., Brennan, P., Mello-Thoms, C., Dung, P., Pietryzk, M., Talanow, R., & Reed, W. (2016). Outcomes knowledge may bias radiological decision-making. *Academic radiology, 23*(6), 760-767.

Louie, T. A., Rajan, M. N., & Sibley, R. E. (2007). Tackling the Monday-morning quarterback: Applications of hindsight bias in decision-making settings. *Social cognition, 25*(1), 32-47.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509-527.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & cognition, 38*(4), 441-451.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174-179.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699 –710.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin, 95*(1), 109-133.

Nestler, S., Blank, H., & Egloff, B. (2010). Hindsight≠ hindsight: Experimentally induced dissociations between hindsight components. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1399-1413.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin & Review, 10*(1), 177-183.

Pezzo, M. V. (2011). Hindsight bias: A primer for motivational researchers. *Social and Personality Psychology Compass, 5*(9), 665-678.

Pohl, R. F. (1992). Der Rückschau-Fehler: Systematische Verfälschung der Erinnerung bei Experten und Novizen [Hindsight bias: Systematic distortion of recollections of experts and novices]. *Kognitionswissenschaft, 3*(1), 38-44.

Pohl, R. F., & Erdfelder, E. (2017). Hindsight bias. In R. F. Pohl (Ed.), *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory* (2nd ed., pp. 424-445). Hove, UK: Psychology Press

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435-451.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological bulletin, 137*(1), 131-148.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science, 7*(5), 411-426.
Sidi, Y., Ackerman, R., & Erez, A. (2018). Feeling happy and (over) confident: the role of positive affect in metacognitive processes. *Cognition and Emotion, 32*(4), 876-884.
Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61-73.
Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553-558.
Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73.
Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & cognition, 46*(8), 1360-1375.
Wallace, H. M., Chang, M., Carroll, P. J., & Grace, J. (2009). I knew it all along, unless I had to work to learn what I know. *Basic and Applied Social Psychology, 31*(1), 32-39.
Werth, L., & Strack, F. (2003). An inferential approach to the knew-it-all-along phenomenon. *Memory, 11*(4-5), 411-419.
Winman, A., Juslin, P., & Björkman, M. (1998). The confidence–hindsight mirror effect in judgment: An accuracy-assessment model for the knew-it-all-along phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 415-431.
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.