Memory & Cognition (2023) 51:1214–1234
https://doi.org/10.3758/s13421-022-01382-3

The "curse of knowledge" when predicting others' knowledge

Jonathan G. Tullis[1] · **Brennen Feder**[1]

Accepted: 10 December 2022 / Published online: 27 December 2022

Abstract
To succeed in a social world, we must be able to accurately estimate what others know. For example, teachers must anticipate student knowledge to plan lessons and communicate effectively. Yet one's own knowledge consistently contaminates estimates about others' knowledge. We examine how one's knowledge influences the calibration and resolution of participants' estimates of novices' knowledge. Across four experiments, participants studied trivia questions and estimated the percentage of novice participants who would know the answer across multiple study/estimation rounds. When participants were required to answer the question before estimating what novices would know, studying the facts impaired both the calibration and resolution of the estimates. Studying the facts reduced the validity of one's experiences for predicting novices' knowledge, and estimators utilized their own experiences less when predicting novices' knowledge as they studied. Experimentally reducing reliance on one's own knowledge did not improve the accuracy of estimates. The results suggest that learning impairs the accuracy of judgments of others' knowledge, not because estimators rely too heavily on their own experiences, but because estimators lack diagnostic cues about others' knowledge.

Keywords Perspective-taking · Metacognition · Curse of knowledge · Expertise bias · Egocentrism

Accurately estimating others' knowledge is crucial to thriving in social environments. Politicians, advertisers, and even scientists who can predict what their audiences will understand can better construct persuasive and comprehensible messages. For example, teachers who successfully predict what their students will know can effectively tailor their pedagogy to support student learning (Sadler et al., 2013), while those who struggle to accurately take the perspective of novices may not communicate effectively (Wieman, 2007). Similarly, doctors need to understand patients' knowledge to effectively convey information about appropriate medication use (Hargis & Castel, 2019). Yet one's own knowledge can significantly bias estimates of others' knowledge across a variety of situations (e.g., Ghrear et al., 2016). Understanding how and why one's own knowledge influences estimates of others' knowledge may enable us to predict systematic errors in estimates and suggest how to improve our estimates (Epley & Waytz, 2009). In four experiments, we examined how and why one's own knowledge influences the accuracy of predictions about others' knowledge.

Our own knowledge biases or contaminates our ability to reason about others across many kinds of social judgments. Estimates of others' mental states are often egocentrically biased (e.g., Birch & Bloom, 2007; Keysar & Barr, 2002). For instance, predictions about others are biased towards one's own knowledge when estimating whether others understand the meaning of idioms (Keysar & Bly, 1995), whether others interpret a message as sarcastic (Keysar, 1994), and whether others know the outcome of historical events (Fischhoff, 1975). When estimating others' knowledge, estimators predict that a greater number of others will know an answer when they know the answer than when they do not, which is often referred to as a "curse of knowledge" (Birch, 2005; Birch & Bloom, 2003; Fussell & Krauss, 1992; Kelley & Jacoby, 1996; Nickerson et al., 1987) or an expertise bias. Biases towards one's own knowledge may explain why teachers systematically overpredict student knowledge across a variety of age levels and domains (Berg & Brouwer, 1991; Friedrichson et al., 2009; Goranson, described in Halim & Meerah, 2002; Kelley, 1999; Sadler et al., 2013). The curse of knowledge bias is widespread; it affects judgments across a wide range of disciplines,

✉ Jonathan G. Tullis
tullis@arizona.edu

[1] Department of Educational Psychology, University of Arizona, 1430 E. Second St., Tucson, AZ 85721, USA

including medicine, law, education, business, and politics (e.g., Hinds, 1999; Keysar, 1994; Keysar & Bly, 1995), and affects reasoning across multiple cultures (Heine & Lehman, 1996; Pohl et al., 2002). The contaminating effects of one's own knowledge when predicting others' knowledge may indicate a weakness in adults' theory of mind, in which we acknowledge that others' mental states are different than our own (e.g., Keysar et al., 2003).

In addition to its pervasiveness, the expertise bias is robust: eliminating or reducing the expertise bias is difficult. The influence of egocentric projections is *not* diminished when participants are explicitly warned to avoid the curse of knowledge (Pohl & Hell, 1996) or when participants are instructed to focus on another's perspective prior to or during perspective taking (Damen et al., 2020). Even when in their best interest to do so, people struggle to ignore their own private knowledge when estimating what others know (Camerer et al., 1989). For example, rewards and punishments that incentivize ignoring one's own knowledge do not reduce egocentric biases in accounting decisions (Kennedy, 1995). Despite its pervasiveness and robustness, we do not fully understand the mechanisms underlying the curse of knowledge or how it changes with learning.

Across four experiments, we examined how repeated exposures to trivia questions bias estimates of others' knowledge. Participants studied trivia questions and predicted how well novices would perform on those questions. We examined two central questions. First, we explicitly tested whether learning across multiple study trials impairs the resolution of judgments of others' knowledge. Prior research and theory have focused almost exclusively on the calibration of estimates; calibration reflects the degree to which a person's average predicted performance corresponds to actual average performance (i.e., whether mean estimates are too high or too low; Hacker et al., 2008). Research consistently shows that people overestimate others' knowledge when they know the correct answer (Nickerson et al., 1987). In contrast, little research has examined the impact of expertise on resolution, which indicates one's ability to decipher between easy and difficult items. Resolution is important because accurately distinguishing between the difficulty of items is vital to make effective study choices, especially under limited time (Kornell & Metcalfe, 2006; Metcalfe & Finn, 2008; Tullis & Benjamin, 2011). For example, accurate resolution of judgments about students' knowledge may be critical for teachers to plan effective lessons (Thiede et al., 2018). Teachers need to know which items are easy and which are difficult to organize their instructional time and activities.

Examining both resolution and calibration of judgments of others' knowledge is important because the bases of these kinds of accuracy are different. Comparisons between individual items (i.e., comparing the difficulty of one target to other targets in the list) likely drive the accuracy of resolution (Susser et al., 2013). In contrast, the overall task structure (e.g., the total number of studied items, the type of test, and the number of practice trials) may drive the accuracy of calibration (Connor et al., 1997; Thiede & Dunlosky, 1994). Because the accuracy of resolution and calibration are based upon different factors, the impact of learning on resolution and calibration may differ.

Second, we measured how and why the resolution of estimates about others' knowledge worsens across multiple rounds of learning. Prior research has focused largely on differences in predictions when an estimator knows the answer compared to when an estimator does not know the answer (e.g., Kelley & Jacoby, 1996; Thomas & Jacoby, 2013; Tullis, 2018), rather than across learning of the ideas within participants. Measuring change across multiple exposures to trivia questions can reveal underlying shifts in learners' use of cues as they gain experience with the questions. Understanding these patterns can show how the development of expertise changes predictions of novices and challenges perspective taking.

To understand why predictions change across rounds, we employed the cue-utilization approach to perspective taking (Koriat, 1997; Tullis, 2018). The cue-utilization approach suggests that people infer others' knowledge from weighing salient cues about others' knowledge. As in prior literature, the cues we examined included one's ability to answer the question and how long it took to answer the question (Tullis, 2018). Using the lens model of metacognition (Bröder & Undorf, 2019), we assessed how the validity of these cues changed with learning (i.e., how their relationship to normative difficulty changed across repetitions) and how the utilization of the cues changed with learning (i.e., how their relationship to judgments of others changed across repetitions). Understanding how the validity and utilization of these cues change can reveal the underlying mechanisms that cause the curse of knowledge. For example, some theories of the curse of knowledge suggest that our judgments about others become inaccurate because we fail to inhibit our own experiences when predicting others' knowledge (Brown-Schmidt & Hanna, 2011). If this inhibition explanation were true, we would expect utilization of one's own experiences to be larger than their validity and we would expect that difference to widen across training. Comprehending how and why repetitions affect estimates of novices' knowledge may ultimately show how we can produce accurate estimates of others' knowledge.

Experiment 1

In this first experiment, participants answered trivia questions, received the correct answer, and estimated the percentage of novice participants that would know the answer. Participants studied the questions and estimated the percentage

of novices who would know the answer three times. Three rounds of learning and estimates allow us to track how the resolution and calibration of judgments change with learning. Further, multiple rounds of learning and judgments allow us to test how the validity and utilization of personal experiences change with learning.

Method

Participants Prior research examining how participants estimate what other people know found Cohen's *d* effect sizes that ranged from 0.38 to over 1 (Tullis, 2018). A power analysis using the G*Power computer program (Faul et al., 2007) indicated that a total sample of 130 participants would be needed to detect the smallest effect size found in related prior literature ($d = 0.38$) with alpha at 0.05 and power of 0.80 using a one-way analysis of variance (ANOVA) with correlation among measures of 0, to be conservative. Ultimately, we collected data from 131 participants, who earned partial course credit for introductory educational psychology courses by participating.

Materials Forty general knowledge trivia questions were selected from existing databases (Nelson & Narens, 1980; Tauber et al., 2013). Questions were selected to encompass a wide range of difficulties and a variety of topics, including geography, entertainment, sports, art, science, and history. The normative difficulty of the questions for this experiment and all subsequent experiments in the manuscript was determined by the ability of a separate sample of 100 participants from the same participant pool to answer these questions. These 100 participants received the questions in a random order, entered their answers, and received no feedback (as described in Tullis, 2018). Normative difficulty of each question was the percent of the 100 participants who correctly answered each question. Answers were only counted as correct, across the prior sample and in the current studies, if the participant spelled the answer correctly. Accepting only correctly spelled answers allows for clear and standardized analyses across this study and prior research (e.g., Tauber et al., 2013). The questions ranged in normative difficulty from 2% to 89% correct, with a mean percentage correct of 44% ($SD = 25\%$).[1]

Procedure Participants completed the experiment on desktop computers in a lab while up to three other participants completed the experiment at the same time. The program was created in MATLAB using the Psychophysics Toolbox (Brainard, 1997) and CogSci Toolbox (Fraundorf et al., 2014). The procedure is displayed in the top section of Fig. 1. For each trivia question, participants first answered the question, were given feedback about whether their answer was correct or incorrect, were told what the correct answer was, and then estimated what percent of other participants would know the correct answer on a scale of 0% to 100%. Instructions included, "You will estimate what percent of other participants will know the answer to each question from 0% (*no one*) to 100% (*everyone*) without having studied it. If you think half of the other participants will answer correctly without studying it, you should say 50% of the other participants."

Participants were warned that they may see trivia questions multiple times, but that they should estimate the percentage of other participants who would be able to answer each question without studying it. After answering and rating all of the 40 trivia questions, the order of the trivia questions was randomized and participants completed a second round of answering, feedback, and estimating what others know that was identical to the first. Finally, participants completed a third round of answering the questions, receiving feedback, and estimating what others know; the third round was identical to the first two rounds. Participants were not explicitly told that they were starting the second or third rounds, but they were always asked to estimate the percentage of other participants who could answer each question without studying it. All aspects of the procedure were self-paced, and participants' response times were recorded.

Analytic procedure

Our primary analyses examined the accuracy of judgments about others' knowledge. More specifically, we examined how the resolution and calibration of judgments of others' knowledge change with learning. In our experiments, resolution describes participants' abilities to decipher between normatively easy and difficult questions and is calculated using Pearson correlations between estimates and normative difficulty for each participant.[2] Larger Pearson correlations indicate better resolution, which means that participants more accurately decipher between which questions are easy

[1] Performance across the questions within this sample was strongly correlated with performance across questions from the sample described in Nelson and Narens (1980), $r = .87$, and from the sample in Tauber et al. (2013), $r = .89$. Further, overall performance levels did not differ between this sample and that in Nelson and Narens (1980), $M = 0.54$, $SD = 0.27$, $t(39) = 1.79$, $p = .08$, $d = 0.29$, but participants in this sample answered more questions correctly than participants in Tauber et al. (2013), $M = 0.35$, $SD = 0.27$, $t(39) = 10.99$, $p < 0.001$, $d = 1.76$.

[2] Resolution in metacognitive research is typically calculated using gamma correlations or signal detection theoretic measures because the predicted variable (e.g., recall) is categorical and dichotomous (1 = recalled, 0 = not recalled). We utilize Pearson's *R* to calculate resolution because the predicted variable ranges from 2% to 89% for each participant and it is based upon a ratio scale.

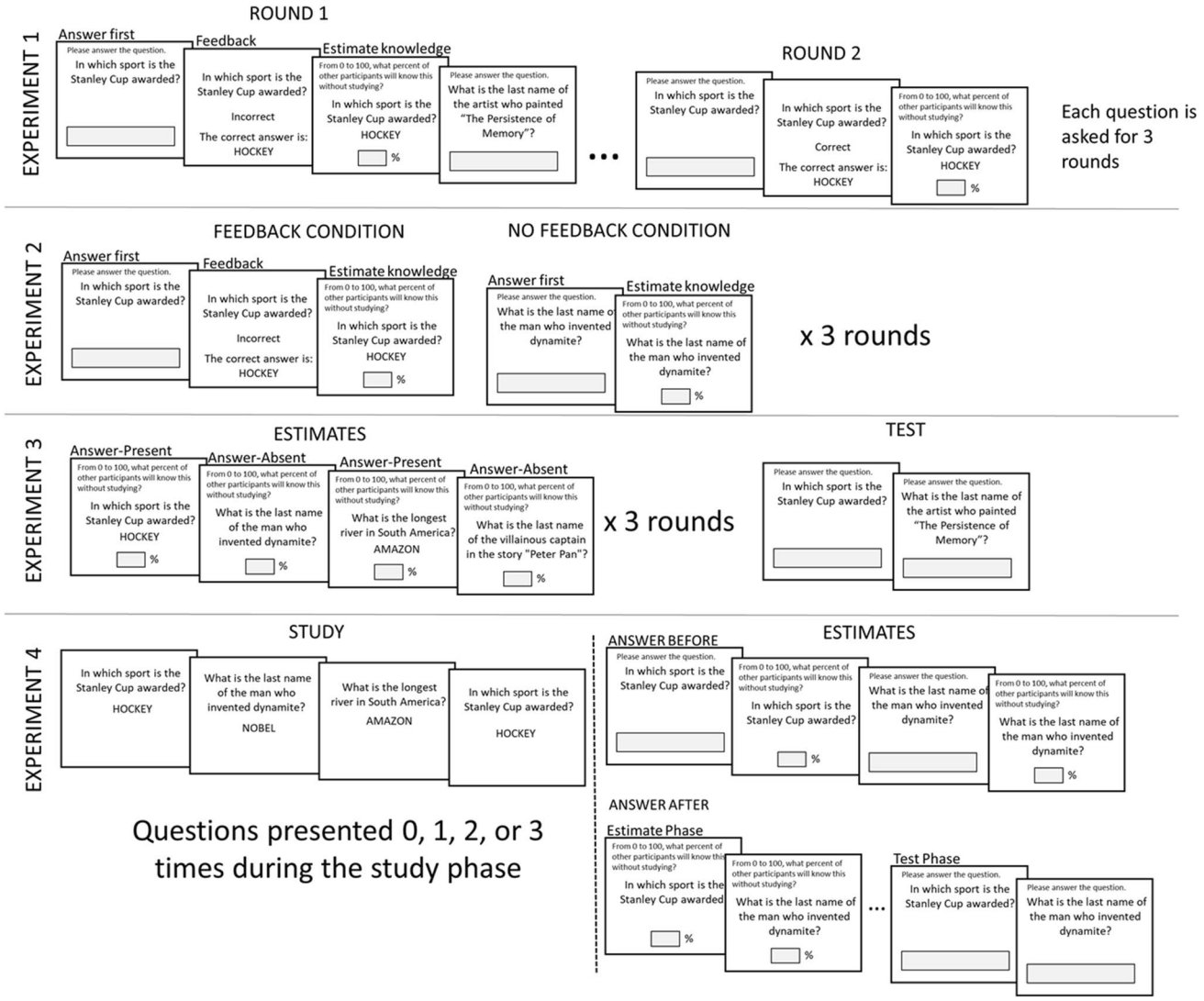

Fig. 1 The general procedures of the four experiments

and difficult for others. Calibration indicates the relationship between the average normative difficulty and the global average of estimates of others' knowledge. We calculate calibration by subtracting the proportion of the tested sample who answered correctly from the proportion estimated to have answered it correctly. Positive calibration scores indicate overestimates of others' knowledge and negative calibration scores indicate underestimates of others' knowledge. Calibration scores closer to zero indicate more accurate overall average assessments of others' knowledge.

We additionally examined why resolution and calibration change during the study. Our experiments allow us to map the utilization (how strongly cues are tied to estimates) and validity (how strongly the cues are tied to normative difficulty) of metamnemonic cues across rounds. Two mnemonic cues have been identified as contributing to predictions about others' knowledge: (1) One's own ability to answer the question and (2) the time it takes to answer each question (Kelley & Jacoby, 1996; Koriat, 2008; Thomas & Jacoby 2013; Tullis, 2018). As shown in Fig. 2, we utilized the lens model of metacognition (Bröder & Undorf, 2019; Brunswick, 1952) to calculate the utilization and validity of metacognitive cues (i.e., one's own accuracy and time needed to answer) for making estimates. *Utilization* shows how strongly metacognitive cues predict judgments of others' knowledge; a linear regression between metacognitive cues and estimates yields a beta weight that represents utilization. *Validity* reveals how strongly those same metacognitive cues predict others' knowledge; a linear regression between metacognitive cues and normative difficulty yields a beta weight that represents validity. In other words, utility describes how strongly participants are using their own metacognitive cues when making estimates, while validity

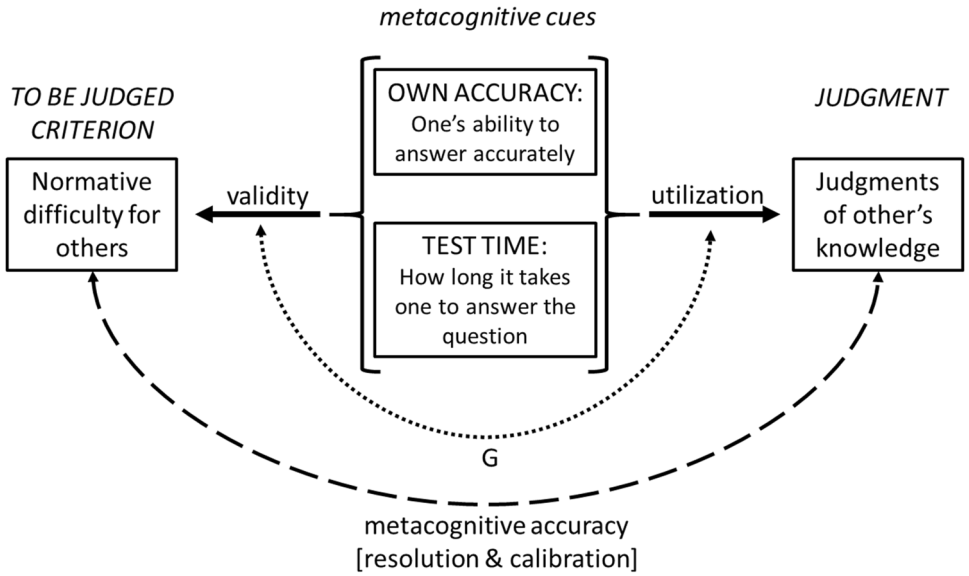

Fig. 2 A schematic representation of the lens model analyses used in our experiments. The left side models the normative difficulty of the trivia questions based upon a linear combination of metacognitive cues (own accuracy and test time) to produce a beta weight called cue validity. The right side models judgments of others' knowledge based upon a linear combination of metacognitive cues to produce a beta weight called cue utilization. G describes the degree of optimal weighting between validity and utilization of metacognitive cues

describes how diagnostic those cues are of normative difficulty. Beta weights representing validity and utilization account for potential cue intercorrelations that individual bivariate correlations cannot. Changes in validity reflect the usefulness of cues in predicting normative difficulty due to changes in one's own learning and do not indicate any change in metacognition. Finally, the lens model also yields a matching index (G), which measures how well an individual judge's cue weighting corresponds to optimal cue weighting. When G is high, the judge is optimally weighing the cues to predict others' knowledge. As greater amounts of noise influence the judge's estimates, G decreases. Specific code and documentation for these analyses are available on the Open Science Framework (https://osf.io/2ngbq/?view_only=ada6614377a24bc797df3046dcee2872). The separation of utilization and validity of metacognitive cues is well suited to our research questions because it can show if changes in judgments across learning are caused by overutilization of metacognitive cues related to one's own knowledge or greater introduction of noise into the judgment process.

Finally, for each of our analyses, we report the Bayes factors to describe the strength of evidence in favor of the alternative hypothesis (BF_{10}). Bayes factors were calculated using the BayesFactor library in R.

Results

Data and analytic code from this and following experiments are available on the Open Science Framework (https://osf.io/2ngbq/?view_only=ada6614377a24bc797df3046dcee2872).

Ability to answer correctly First, we examined whether participants' ability to answer the trivia questions changed across round and the means are displayed in the top row of Table 1. A one-way repeated-measures ANOVA on answer accuracy revealed a significant effect of round on accuracy, $F(2, 260) = 1428.59$, $p < .001$, $\eta_p^2 = .917$, $BF_{10} = 1.25E139$. In other words, participants learned the correct answers to the trivia questions across rounds.

Metacognitive accuracy Second, we examined whether metacognitive accuracy changed across rounds. We first examined the calibration of judgments (i.e., the proportion prediction minus the proportion of novices who correctly answered the question) across rounds, as shown in the fourth row of Table 1. A one-way ANOVA on calibration showed a significant effect of round, $F(2, 260) = 8.31$, $p < .001$, $\eta_p^2 = .060$, $BF_{10} = 54.04$. A specific post hoc paired t test showed that calibration was significantly worse in Round 3 than in Round 1, $t(130) = 3.21$, $p = .002$, Cohen's $d = 0.28$,

Table 1 Proportion of questions answered correctly (top row), time needed to answer each question (second row), mean estimate of the proportion of others who could answer each question correctly (third row), mean calibration (proportion estimate minus normative proportion correct; fourth row), mean resolution (Pearson correlation between estimate and normative difficulty; fifth row), and optimal weighting of cues (G; bottom row) across three rounds in Experiment 1 (standard deviations are shown in parentheses)

Dependent variable	Round 1	Round 2	Round 3
Accuracy	.43 (.16)	.81 (.16)	.88 (.13)
Test time	10.60 (4.79)	4.85 (1.67)	4.17 (1.38)
Estimates	.53 (.12)	.55 (.15)	.57 (.16)
Calibration	.09 (.12)	.11 (.15)	.12 (.16)
Resolution (r)	.55 (.15)	.51 (.16)	.49 (.18)
Optimal weighting (G)	.94 (.10)	.83 (.36)	.65 (.57)

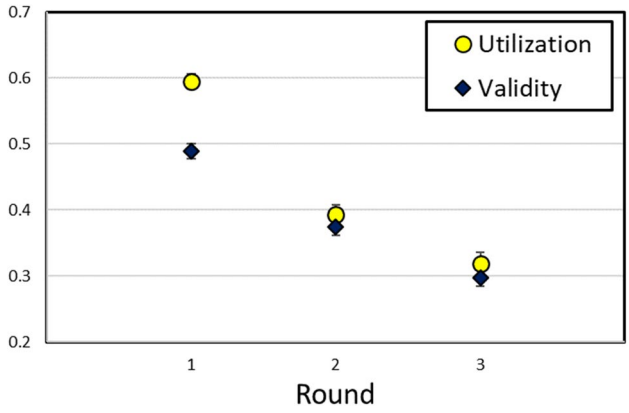

Fig. 3 The beta weights from the lens model across rounds in Experiment 1. Error bars show standard errors of the mean

$BF_{10} = 12.36$. Similarly, we examined how the resolution[3] of judgments, as indicated by the Pearson correlation between estimates and normative difficulty, varied across rounds. A one-way repeated-measures ANOVA on resolution showed a significant impairment in resolution across rounds, $F(2, 260) = 9.84$, $p < .001$, $\eta_p^2 = .070$, $BF_{10} = 205.68$. A specific post hoc comparison between the first and third round showed a significant decrease in resolution, $t(130) = 3.88$, $p < .001$, Cohen's $d = 0.34$, $BF_{10} = 104.37$.

Utilization and validity of cues The utilization and validity beta weights are shown in Fig. 3. A 2 (type of measure: utilization vs. validity) × 3 (round) repeated-measures ANOVA on the weights resulting from the regression analysis revealed a significant interaction, $F(2, 260) = 14.43$, $p < .001$, $\eta_p^2 = .10$, $BF_{10} = 68.61$, a significant effect of round, $F(2, 260) = 139.77$, $p < .001$, $\eta_p^2 = .52$, $BF_{10} = 1.09E61$, and a significant effect of type of measure, $F(1, 130) = 22.13$, $p < .001$, $\eta_p^2 = .15$, $BF_{10} = 114.32$. The interaction shows that estimators' utilization of metacognitive cues related to their experience was too large in the first round, but decreased to a greater extent than validity across rounds. Participants show no systematic propensity to overweight metacognitive cues related to their own experiences as they learned the answers. Finally, as shown in the bottom row of Table 1, a one-way repeated-measures ANOVA on G (the degree of match between utilization and validity) shows a significant decrease across rounds, $F(2, 260) = 19.09$, $p < .001$, $\eta_p^2 = .128$, $BF_{10} = 2.18E6$. A specific post hoc t test indicated that G decreased from Round 1 to Round 3, $t(130) = 5.77$, $p < .001$, Cohen's $d = 0.51$, $BF_{10} = 204{,}012$. This indicates that the optimal weighting of metacognitive cues dropped (implying increased noise in judgments) with increased learning across rounds.

Discussion

Both the calibration and resolution of judgments of others' knowledge worsened as participants gained experience with trivia answers. As participants' knowledge of the trivia answers grew, their estimates of others' knowledge increased, and participants produced greater overestimates of others' knowledge. Further, participants' ability to distinguish between easy and difficult items degraded across rounds. Growing overestimates of others' knowledge with learning mimics the research on the curse of knowledge (e.g., Birch, 2005; Birch & Bloom, 2003; Fussell & Krauss, 1992; Kelley & Jacoby, 1996; Nickerson et al., 1987). The impairment of resolution of predictions with experience contradicts prior research, which suggests that greater knowledge can improve resolution (Kelley & Jacoby, 1996; Thomas & Jacoby, 2013). However, prior research has compared resolution between a group that knows the answer and a group that does not. In this experiment, all participants saw the correct answer and we tracked changes in resolution with additional exposures to the correct answer. Knowing the correct answer can provide important metacognitive cues about the normative difficulty of a question (e.g., the familiarity of the *answer*), which may increase metacognitive resolution. Additional exposures beyond initial learning of the answer in this experiment, however, led to impairments in resolution.

Decrements in resolution correspond to a decrease in the validity of cues across rounds. One's own ability to answer and the time it takes to answer each question, both of which are salient metacognitive cues, become less tied to normative difficulty across rounds. Reductions in validity do not

[3] Resolution is typically used to describe the correlation between one's own predictions of one's memory with one's actual memory. Here, we use resolution to indicate the correlation between estimates of others' knowledge and others' knowledge.

indicate anything about one's metacognition, but show that one's own experiences are objectively less reflective of novices' knowledge. As the cues became less valid for predicting others' knowledge, participants utilized these cues less. The results suggest that participants did not rely too heavily on their own experiences when estimating what others know across repetitions. Instead, participants' utilization of their own experiences dropped more than the validity of those cues dropped. Reducing the utilization of cues related to one's own experiences, however, introduced greater noise into participants' estimates. As estimators shifted away from their own experiences, they did not shift towards using more valid cues. Reducing the availability and salience of valid cues reduces the accuracy of estimates of others' knowledge. In addition to the reduction in the utilization of cues related to one's own experiences, participants' optimal weighting of their own cues decreased. Estimators were not utilizing their own experiences as precisely as they could have. Greater noise in using their own cues and reductions in valid cues likely impair both the calibration and resolution of estimates of others' knowledge. In summary, two factors likely underly impairments in the accuracy of judgments about others. First, participants utilized their own cues less across rounds (because the validity of these cues dropped), which can allow other non-valid cues to influence judgments. Second, participants mis-weighted their own cues across rounds. We replicated and extended these results in Experiment 2.

Experiment 2

Experiment 1 showed that the resolution and calibration of predictions about others became less accurate as estimators gained practice with the questions. Yet the results of Experiment 1 are correlational: Learning and less accurate metacognitive predictions both happened across rounds. In Experiment 2, we experimentally manipulated learning across rounds to isolate the impact of increased knowledge on predictions. To experimentally manipulate learning across rounds, participants were provided feedback about the correct answer only for a random half of the questions; for the other half, participants were never provided feedback about the correct answer. We predict that the validity of cues for questions with feedback will decrease across rounds as in Experiment 1 because the feedback introduces noise in these cues. However, for questions without feedback, the validity of cues should remain consistent across rounds because no new knowledge is introduced that would contaminate those cues. This will allow us to test whether reductions in the validity of one's own cues are necessary to reduce the accuracy of judgments about others across rounds. As in Experiment 1, this experiment aimed to test two primary research questions: (1) Does learning impair the resolution of participants' estimates of others' knowledge? and (2) How do utilization and validity of metacognitive cues about others' knowledge change with learning?

Participants

We aimed to match the same sample size as Experiment 1 and collected a total sample of 132 participants from introductory educational psychology classes at the University of Arizona. One participant was ultimately excluded because they supplied the same estimate (0%) for all items, which precludes meaningful calculations of resolution and the lens model.

Materials

The same 40 trivia questions used in the prior experiment were utilized here.

Procedure

The experiment proceeded similarly to Experiment 1 with one change: Half of the questions were randomly assigned to the no-feedback condition and half to the feedback condition within participants. The procedure is depicted in the second row of Fig. 1. For questions in the no-feedback condition, participants answered the trivia question and then immediately estimated what percent of others would know the answer without studying it. Participants never saw the correct answers for questions in this condition. For questions in the feedback condition, participants answered the question, received feedback about whether their answer was correct or incorrect while seeing the correct answer, and finally estimated the percentage of other participants who would know the answer without studying it. As in Experiment 1, participants completed three rounds of study and prediction and always estimated the percentage of others who would know the answer without studying it.

Results

Ability to answer correctly To test whether participants learned the correct answers across repeated presentations and whether that gain in knowledge depended upon feedback, we first tested participants' accuracy on the trivia questions as a function of round and feedback condition. A 2 (feedback condition) × 3 (round) repeated-measures ANOVA showed a significant interaction between conditions, $F(2, 260) = 1028.16$, $p < .001$, $\eta_p^2 = .89$, $BF_{10} = 5.39E136$, a significant main effect of feedback, $F(1, 130) = 884.56$, $p < .001$, $\eta_p^2 = .87$, $BF_{10} = 5.38E88$, and a significant effect of round, $F(2, 260) = 1146.14$, $p < .001$, $\eta_p^2 = .90$, $BF_{10} = 1.20E31$. As shown in Table 2, participants

Table 2 Proportion of questions answered correctly (top half) and time required to answer questions (bottom half) in feedback and no-feedback conditions (standard deviations are shown in parentheses)

	Round 1	Round 2	Round 3
	Proportion correct		
Feedback cond	.40(.18)	.79 (.16)	.88 (.13)
No-feedback cond	.37 (.17)	.38 (.18)	.37 (.18)
	Test time		
Feedback cond	10.72 (3.41)	4.94 (1.54)	4.21 (1.25)
No-feedback cond	11.16 (4.31)	5.71 (1.58)	4.51(1.23)

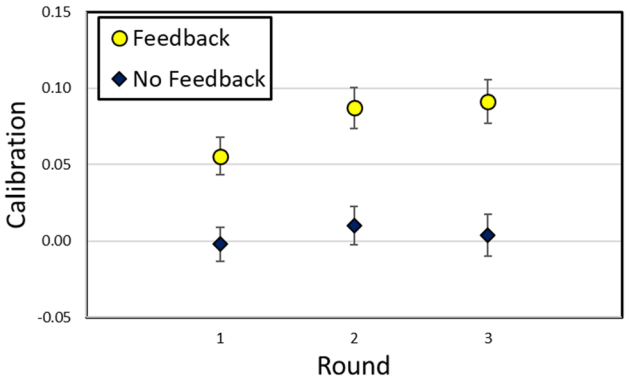

Fig. 4 The calibration (proportion predictions minus proportion normative difficulty) by feedback condition and round in Experiment 2. Error bars show standard errors of the mean

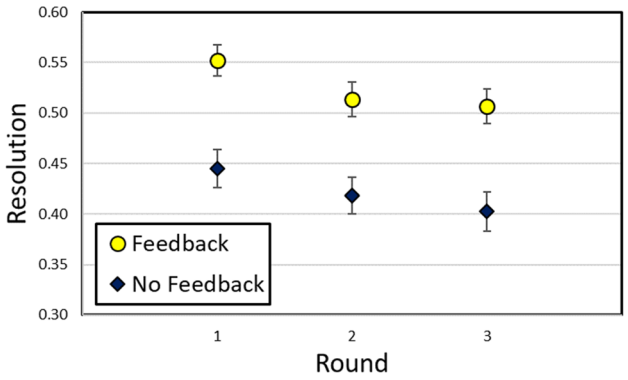

Fig. 5 The resolution (Pearson correlations between predictions and normative difficulty) of knowledge estimates by feedback condition and round in Experiment 2. Error bars show standard errors of the mean

became more accurate across rounds when they were provided feedback, but their accuracy did not change across rounds when no feedback was provided.

Metacognitive accuracy Next, we examined the calibration of estimates (proportion prediction minus the proportion of others who correctly answered the question) across repetitions, feedback condition, and their interaction, which is shown in Fig. 4. The 2 (feedback condition) × 3 (round) ANOVA on calibration revealed a significant interaction, $F(2, 262) = 8.20$, $p < .001$, $\eta_p^2 = .059$, $BF_{10} = .47$, a significant main effect of feedback, $F(1, 130) = 60.06$, $p < .001$, $\eta_p^2 = .316$, $BF_{10} = 3.70E28$, and a significant main effect of round, $F(2, 260) = 10.18$, $p < .001$, $\eta_p^2 = .073$, $BF_{10} = 7.95$. Specific post hoc comparisons between Round 1 versus Round 3 showed that calibration became worse in the feedback condition, $t(130) = 3.74$, $p < .001$, Cohen's $d = 0.33$, $BF_{10} = 64.76$, but did not significantly change in the no-feedback condition, $t(130) = 1.12$, $p = .27$, Cohen's $d = 0.10$, $BF_{10} = 0.18$.

Next, we examined the resolution of metacognitive predictions across rounds, which is shown in Fig. 5. The 2 (feedback condition) × 3 (round) ANOVA on resolution revealed a significant main effect of feedback, $F(1, 130) = 22.68$, $p < .001$, $\eta_p^2 = .149$, $BF_{10} = 7.28E10$, and a significant main effect of round, $F(2, 260) = 4.52$, $p = .012$, $\eta_p^2 = .034$, $BF_{10} = .20$. The interaction between feedback and round did not reach significance, $F(2, 260) = .18$, $p = .84$, $\eta_p^2 = .001$, $BF_{10} = 0.03$. Specific post hoc comparisons showed that resolution worsened from Round 1 to Round 3 in the feedback condition, $t(130) = 2.16$, $p = .03$, Cohen's $d = 0.19$, $BF_{10} = 0.92$, and in the no-feedback condition, $t(130) = 2.11$, $p = .04$, Cohen's $d = 0.19$, $BF_{10} = 0.83$.

Utilization and validity of cues As in Experiment 1, we compared how utilization and validity changed across rounds, which is shown in Fig. 6. A 2 (type of measure: utilization vs. validity) × 3 (round) × 2 (feedback condition) ANOVA on correlations revealed a nonsignificant three-way interaction, $F(2, 260) = 2.22$, $p = .11$, $\eta_p^2 = .017$, $BF_{10} = 0.07$. The ANOVA revealed significant interactions between round and measure, $F(2, 260) = 4.66$, $p = .01$, $\eta_p^2 = .035$, $BF_{10} = .10$, and between feedback and round, $F(2, 260) = 59.42$, $p < .001$, $\eta_p^2 = .314$, $BF_{10} = 1.09E19$. The interaction between feedback and measure did not reach significance, $F(1, 130) = 0.07$, $p = .80$, $\eta_p^2 = .001$, $BF_{10} = 0.09$. The ANOVA revealed significant main effects of feedback, $F(1, 130) = 22.04$, $p < .001$, $\eta_p^2 = .15$, $BF_{10} = 2.08E12$, round, $F(2, 260) = 73.58$, $p < .001$, $\eta_p^2 = .361$, $BF_{10} = 5.06E19$, and measure, $F(1, 130) = 43.22$, $p < .001$, $\eta_p^2 = .250$, $BF_{10} = 4.10E7$.

We also computed a 3 (round) × 2 (feedback) repeated-measures ANOVA on G, and the results showed a significant interaction, $F(2, 260) = 3.49$, $p = .032$, $\eta_p^2 = .026$, $BF_{10} = .58$, a main effective of feedback, $F(1, 130) = 7.05$, $p = .009$, $\eta_p^2 = .051$, $BF_{10} = 2.34$, and a main effect of round, $F(2, 260) = 13.96$, $p < .001$, $\eta_p^2 = .097$, $BF_{10} = 18{,}844.47$. Specific post hoc comparisons between the first and third rounds showed significant drops in G for both the feedback

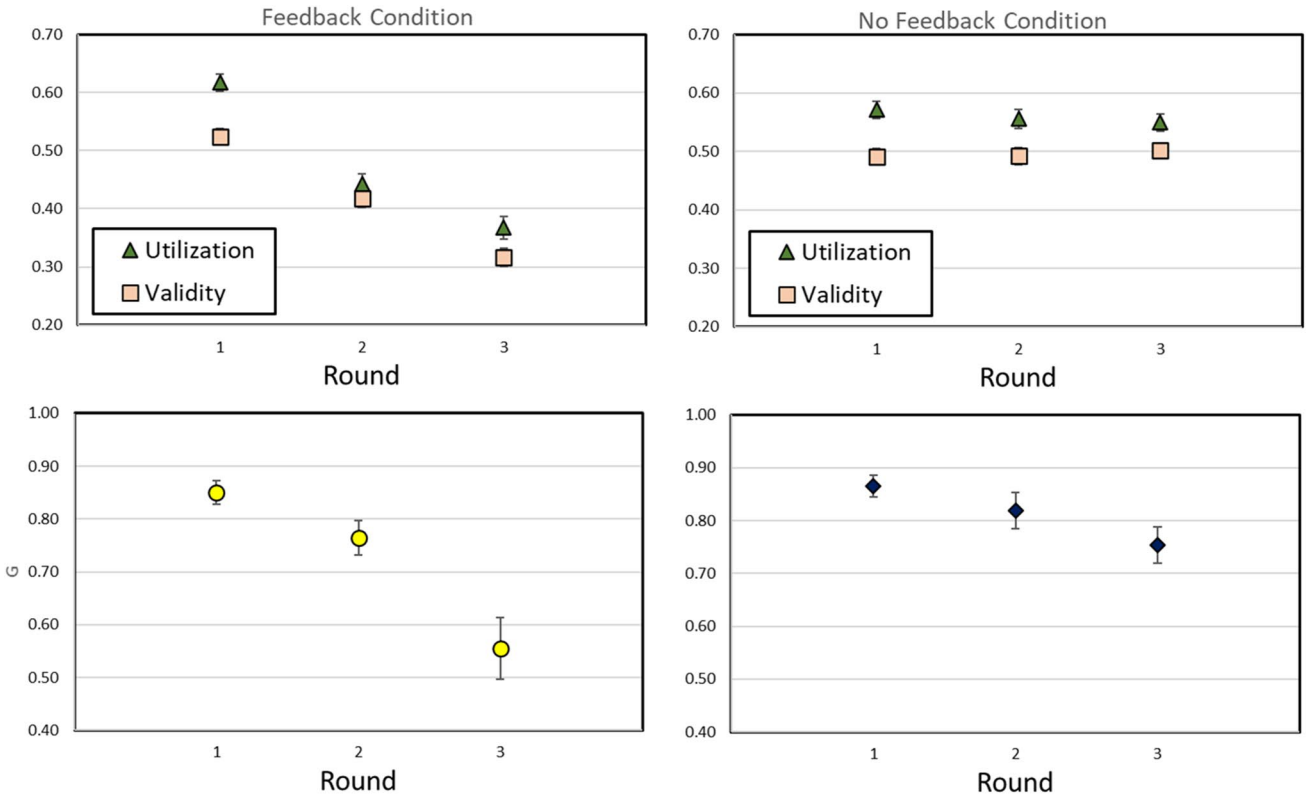

Fig. 6 The beta weights (top row) and G (bottom row) from the lens model across rounds in Experiment 2. Error bars show standard errors of the mean

group, $t(130) = 3.96$, $p < .001$, Cohen's $d = 0.35$, $BF_{10} = 138.17$, and the no-feedback group, $t(130) = 2.38$, $p = .02$, Cohen's $d = 0.21$, $BF_{10} = 1.46$.

Discussion

In Experiment 2, we experimentally manipulated whether participants learned across rounds by providing feedback to only half of the trivia questions. The results of the feedback condition very closely replicated Experiment 1. Providing feedback yielded learning, as participants' answers improved within the feedback condition. With increased learning, calibration became worse across rounds. More specifically, for questions in the feedback condition, participants predicted that more people would know the answer than did and overpredictions grew across rounds. Overpredictions for questions with feedback corroborate prior research (Tullis, 2018). When estimators see the correct answer, they may exhibit hindsight bias, in which they think they should have known an answer, and this bias may affect their estimates of what others know. In contrast, estimates did not increase when feedback was withheld, so calibration remained consistent across rounds in the no-feedback condition. These results suggest that overall predictions are sensitive to the amount of knowledge that one has. As one gains more knowledge, mean predictions of others' knowledge increase.

As in Experiment 1, resolution became significantly less accurate across rounds. Notably, and in contrast to calibration, additional exposures impaired resolution in both the feedback and no-feedback conditions. Repeated questioning decreased participants' resolution whether they received feedback or not, indicating that participants' estimates become contaminated with greater noise across rounds, even without feedback. The patterns of data between the feedback conditions reveal a dissociation between the accuracy of calibration and resolution. Calibration may be driven by overall level of knowledge, while resolution can be affected by idiosyncratic personal experiences with the facts that distort how well participants utilize their metacognitive cues. These results highlight the importance of measuring both calibration and resolution when assessing metacognition about others' knowledge because the patterns between the two constructs can differ.

Feedback impaired the validity and utilization of cues, but the validity and utilization of cues remained consistent across rounds when feedback was withheld. The validity of cues for questions with feedback dropped significantly across rounds as repeatedly learning the correct answers made one's own experiences less predictive of others'

knowledge. In contrast, when no feedback was provided, one's own experiences remained valid for predicting others' knowledge across repeated exposures to the questions. The patterns for utilization of cues matched the patterns found with validity; utilization of one's own experiences dropped in the feedback condition but remained consistent across rounds when no feedback was provided. No evidence suggests that estimators overutilized their own experiences with learning; in fact, the interaction between round and measures suggests that estimators overutilized their own experiences during early rounds, but this overutilization decreased across rounds. The results also show that noise in weighing one's own cues increased across both conditions (i.e., G decreased). Greater noise combined with a lack of diagnostic cues about others, rather than overutilization of one's own experiences, causes decrements to the accuracy of judgments of others' knowledge during learning.

Finally, the resolution of predictions in the feedback condition was greater than that in the no-feedback condition. Estimates in the Feedback condition more accurately reflected normative difficulty than those in the no-feedback condition, which replicates prior research (Kelley & Jacoby, 1996; Tullis, 2018). The correct answer reveals diagnostic cues about the difficulty of the questions (e.g., the familiarity of the answer itself and the strength of the relationship between the question and answer), which are absent in the no-feedback condition.

Experiment 3

Prior research suggests that the curse of knowledge bias is difficult, if not impossible, to avoid (e.g., Camerer et al., 1989). The cue-utilization framework of knowledge estimation suggests that reliance on one's own knowledge to estimate what others know can be increased or decreased through the judgment conditions (Tullis, 2018). More specifically, when estimators are required to answer the question before estimating others' knowledge, their own ability to answer impacts their estimates more than when they are not required to do so. Requiring learners to answer each question first makes one's ability to answer the question salient and increases the estimators' reliance on their own experience (Tullis, 2018). In Experiments 1 and 2, learners were required to answer each question before estimating the percent of others that could answer the question. In Experiment 3, we examined how the calibration and resolution of judgments of others' knowledge change across learning when estimators are not required to answer each question first. When learners are not required to answer the question first, metacognitive cues related to one's own experiences are less salient, and estimators may be less likely to utilize their own cues. Consequently, reducing the requirement to answer before estimating others' knowledge may reduce the impact of one's knowledge on judgments of others' knowledge.

Method

Participants As in Experiment 1, 131 students in introductory educational psychology courses participated in partial fulfillments of their course requirements. One participant was dropped because they estimated 90% of others would know every answer, which precludes meaningful calculation of resolution and lens model statistics.

Materials The same 40 trivia questions used in Experiment 1 were utilized here.

Procedure As shown in the third row of Fig. 1, the experiment proceeded similarly to Experiment 2, with one significant change. Participants did not input the answer for each trivia question before estimating how many other participants would know the answer. In the answer-present condition, the answer was provided on the screen at the same time as the question, and participants provided their estimates when both question and answer were displayed. In the answer-absent condition, participants saw the questions but were never provided with the answers, and participants provided their estimates with just the question present. As in Experiment 2, questions were randomly assigned to condition. Participants took one final trivia test after completing all three rounds of estimation so that we could assess their final knowledge. During the final test, each trivia question was displayed one at a time in a new random order and participants attempted to answer each one. The final test was included to measure whether participants learned with feedback.

Results

Ability to answer correctly We first assessed participants' accuracy on the final test as a function of feedback condition. Accuracy on the final test was higher for questions in the answer-present condition ($M = .76$, $SD = .19$) than for those in the answer-absent condition ($M = .33$, $SD = .17$), $t(129) = 33.65$, $p < .001$, $d = 2.96$. $BF_{10} = 1.15E62$.

Metacognitive accuracy We examined the calibration of estimates across rounds and by answer condition, as shown in Fig. 7. A 2 (answer present vs. answer absent) × 3 (round) repeated-measures ANOVA on calibration showed a significant effect of the presence of the answer, $F(1, 129) = 54.55$, $p < .001$, $\eta_p^2 = .30$, $BF_{10} = 1.74E31$. Neither the interaction, $F(2, 258) = 2.28$, $p = .10$, $\eta_p^2 = .017$, $BF_{10} = 0.04$, nor the main effect of round, $F(2, 258) = 2.36$, $p = .10$, $\eta_p^2 = .018$, $BF_{10} = .06$, reached significance. The impact of round is

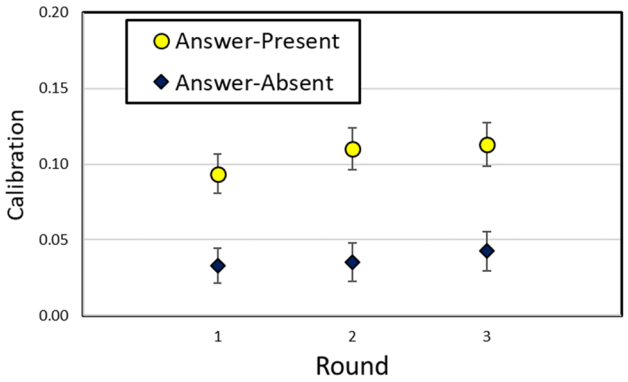

Fig. 7 Calibration by feedback condition and round in Experiment 3. Error bars show standard errors of the mean

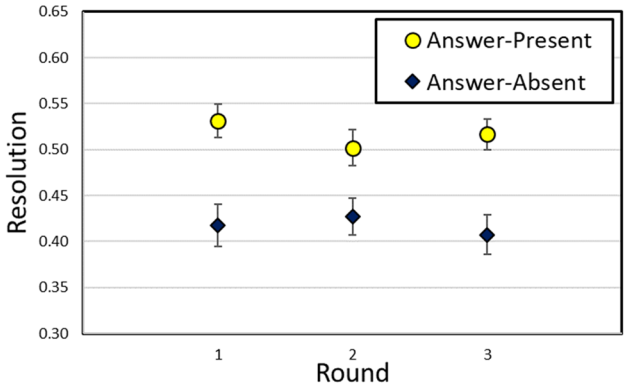

Fig. 8 The resolution of estimates by feedback condition and round in Experiment 3. Error bars show standard errors of the mean

noticeably reduced from Experiment 2 and the Bayes factors suggest strong evidence against its impact.

Next, we examined the resolution of participants' predictions, as shown in Fig. 8. The repeated-measures ANOVA on resolution of judgments showed only a significant effect of answer condition, $F(1, 129) = 23.99$, $p < .001$, $\eta_p^2 = .157$, $BF_{10} = 3.95E12$. Neither the round, $F(2, 258) = .78$, $p = .46$, $\eta_p^2 = .006$, $BF_{10} = 0.02$, nor the interaction of round with the answer condition, $F(2, 258) = 2.13$, $p = .12$, $\eta_p^2 = .016$, $BF_{10} = 0.07$, reached significance.

Utilization and validity of cues We cannot measure the utilization and validity of cues across rounds because participants did not answer each question across round; we have no measure of their ability to answer or their test time except for on the final test. We compared how utilization and validity of cues during the final round of estimates changed with answer conditions, as shown in Table 3. A 2 (type of measure: utilization vs. validity) × 2 (answer present vs. answer absent) repeated-measures ANOVA revealed a significant effect of answer condition, $F(1, 129) = 24.00$, $p < .001$, $\eta_p^2 = .157$, $BF_{10} = 6.96E7$. Neither the interaction, $F(1, 129) = 3.82$, $p = .053$, $\eta_p^2 = .029$, $BF_{10} = 0.37$, nor the main effect of type of measure, $F(1, 129) = 0.68$, $p = .41$, $\eta_p^2 = .005$, $BF_{10} = 0.12$, reached significance. The table shows that the presence of the answer reduced the validity of one's metacognitive cues but also caused estimators to utilize those cues to a lesser degree. Finally, a repeated measures *t* test showed a significant effect of answer condition on the degree of optimal weighting of cues (G), $t(129) = 2.01$, $p = .046$, $d = .18$, $BF_{10} = .69$. This result suggests that, when the answer was present during learning, participants' utilization of metacognitive cues may have matched their validity less than when the answer was not present, which replicates data from Experiment 2.

Table 3 Utilization, validity, and degree of optimal weighting in the answer-present and answer-absent conditions (standard deviations are shown in parentheses)

	Utilization	Validity	G
Answer-present cond	.40 (.17)	.43 (.17)	.71 (.41)
Answer-absent cond	.52 (.20)	.51 (.19)	.80 (.40)

Discussion

This experiment differed from the prior experiments because participants were not required to answer each question before estimating how many others would know it. Removing the requirement reduced the impact of rounds on judgments. Even though participants learned the correct answers across rounds in the feedback condition, neither calibration nor resolution changed across rounds. The stability of calibration and resolution, even in the feedback condition, indicates that estimates of others were only slightly impacted by participants' growing knowledge, in contrast to the prior two experiments.

Broadly, the framing of metacognitive questions and the presence of contrasting conditions can increase the salience of some cues and significantly alter metacognitive judgments (Koriat et al., 2004). Removing the requirement to answer each question likely reduced the salience and reliance on one's own experiences answering each question. Consequently, changes in one's knowledge across rounds did not impact estimates of others' knowledge and produced different patterns of data from the prior two experiments. The data show that one's own ability to answer each question (and its growth across study repetitions) does not automatically impact estimates of others' knowledge. Judgment conditions may modulate the impact of one's own knowledge on estimates of others' knowledge, even across multiple rounds of learning. Reducing the salience of one's own ability to answer a question may be one

means to avoid the curse of knowledge bias when estimating what others know. Yet these conclusions are driven by between-experiment comparisons; between-experiment comparisons should be interpreted cautiously because they may reflect systematic differences in conditions or other nonexperimental differences. In the final experiment, we directly compare answer-required and no-answer-required conditions to test these hypotheses.

Experiment 4

Experiment 4 aimed to answer two central questions: First, does requiring the participant to answer before estimating others' answers tie estimates more strongly to their own knowledge? Second, does anchoring in one's own knowledge cause inaccuracies in estimates of others' knowledge? To answer those questions, we directly compared a condition in which participants answered trivia questions immediately before estimating others' knowledge with a condition in which participants were not required to answer before estimating others' knowledge. If requiring people to answer makes one's own experiences more salient, the impact of one's own knowledge should be more apparent when estimators answer the trivia question before predicting others' knowledge than when estimators are not required to do so. In other words, given the comparisons between Experiments 2 and 3, we expect that requiring participants to answer each question before estimating others' knowledge will increase the utilization of their own experiences.

If inaccuracies in judgments about others' knowledge are caused by overutilization of one's own knowledge and experience (e.g., Keysar et al., 2000), reducing utilization of those cues should improve the resolution of those judgments. We introduced variability in participants' knowledge by exposing them to the trivia questions for different amounts of study repetitions. Some answers were not studied during the initial study round while others were studied up to three times. Variation in study exposures may better mimic natural idiosyncrasies in our knowledge and should make one's own knowledge less reflective of normative difficulty. Learners often utilize natural idiosyncrasies in processing to-be-learned material to accurately predict their own future memory (e.g., Koriat, 1997, 2008; Lovelace, 1984), but individual idiosyncrasies in processing may contribute to inaccurate judgments about others' knowledge. If the curse of knowledge is caused by relying upon one's knowledge too extensively, reducing reliance on those idiosyncratic experiences by not mandating participants answer the questions themselves should improve accuracy of judgments.

Participants

In order to detect a small (Cohen's $d = .30$) effect with power of 0.8 and alpha of 0.05, we collected data from 352 participants in introductory educational psychology classes, who completed the experiment in exchange for partial course credit. We started collecting participants in the lab, as in prior experiments; however, after collecting 132 participants' data in the lab, we shifted data collection online due to the COVID-19 pandemic. The in-person participants completed the experiment in MATLAB and were alternatively assigned to conditions; this resulted in 66 in-person participants in each condition. The online version of the experiment was programmed using Pavlovia and participants were randomly assigned to condition; this resulted in 113 online participants in the answer-before condition and 107 online participants in the answer-after condition. Two online participants in the answer-after condition were excluded because they did not finish the study phase. In total, 179 participants completed the answer-before condition and 173 participants completed the answer-after condition.

Materials

The same 40 trivia questions used in Experiments 1–3 were utilized here.

Procedure

Trivia questions were assigned to four different repetition conditions. Ten questions were never studied, 10 were studied once, 10 were studied twice, and 10 were studied three times, for a total of 60 presentations. The difficulty of the trivia questions across repetition conditions was constrained so that each repetition condition contained approximately equivalently difficult items. To do so, trivia questions were split into 10 different groups based upon their difficulty (i.e., the four easiest were in a group, the next four easiest were in a group). One question from each difficulty group was randomly placed into each repetition condition.

Participants were told "In the first portion of this experiment, you will just study a list of trivia facts presented to you. You will see 60 trivia facts presented in a row. It is possible that sometimes you will see trivia questions a couple of times. Do your best to learn these answers." The trivia questions and answers were presented on the screen one at a time in black 30-point Arial font for 6 seconds each. The 60 presentations of the trivia questions were presented in an entirely random order.

After studying the trivia questions, participants were told that they would judge what percent of other participants would be able to answer those trivia questions without having studied them. Participants in the answer-before condition

Table 4 The proportion answered correctly and test time by condition and study repetitions in Experiment 4 (standard deviations are shown in the parentheses)

	Study repetitions			
	0	1	2	3
	Proportion correct			
Answer before	.32 (.24)	.69 (.23)	.77 (.23)	.83 (.20)
Answer after	.35 (.25)	.67 (.24)	.76 (.22)	.82 (.20)
	Test time			
Answer before	13.8 (8.5)	9.9 (7.6)	7.8 (4.7)	7.9 (6.3)
Answer after	10.5 (8.6)	7.6 (5.3)	7.1 (5.8)	7.0 (5.4)

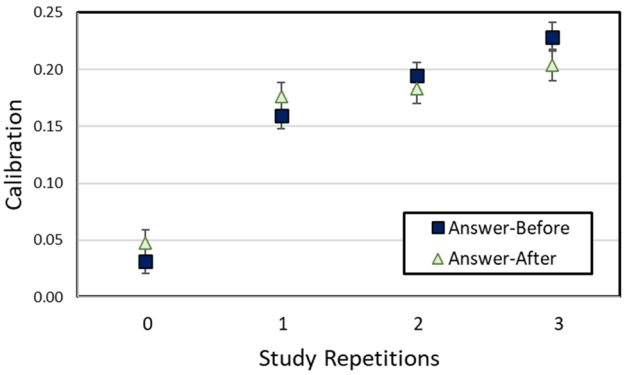

Fig. 9 The calibration of estimates by answer condition and study repetitions in Experiment 4. Error bars show standard errors of the mean

answered each trivia question without corrective feedback and subsequently provided their estimate of what percent of other participants would know the correct answer on a scale of 0% to 100%. Participants in the answer-after condition first estimated the difficulty for all 40 trivia questions on the 0% to 100% scale. After rating the trivia questions, the participants went through the entire list of questions again and provided their best answer to each question (as in Experiment 3).

Results

No differences in data patterns were detected between the in-person and online participants, so the data are combined across all analyses. Separate examinations of in-person and online participants are presented on the OSF webpage as supplemental analyses.

Ability to answer correctly To test whether participants learned with repetitions, we conducted a 4 (repetitions) × 2 (answer condition) mixed ANOVA on proportion answered correctly, as shown in Table 4. The results showed a significant effect of repetitions, $F(3, 1050) = 784.34$, $p < .001$, $\eta_p^2 = .69$, $BF_{10} = 8.95E262$. Neither the interaction between repetitions and answer condition, $F(3, 1050) = 2.18$, $p = .09$, $\eta_p^2 = .006$, $BF_{10} = 0.10$, nor the main effect of answer condition, $F(1, 350) = 0.006$, $p = .94$, $\eta_p^2 < .001$, $BF_{10} = 0.08$, was significant.

Metacognitive accuracy We calculated the calibration of judgments of others' knowledge by the number of study repetitions and the results are shown in Fig. 9. A 4 (study repetitions) × 2 (answer condition) mixed ANOVA on calibration showed a significant interaction, $F(3, 1050) = 4.27$, $p = .005$, $\eta_p^2 = .012$, $BF_{10} = 1.79$, a significant main effect of repetitions, $F(3, 1050) = 245.50$, $p < .001$, $\eta_p^2 = .41$, $BF_{10} = 5.19E115$, but no significant main effect of answering condition, $F(1, 350) = .003$, $p = .96$, $\eta_p^2 < .001$, $BF_{10} = 0.14$. The answer-before condition showed a large decrement to calibration between zero and three repetitions, $t(172) = 14.96$, $p < .001$, $d = 1.13$, $BF_{10} = 4.35E29$. The answer-after condition also showed a large decrement to calibration between zero and three repetitions, $t(178) = 12.59$, $p < .001$, $d = 0.94$, $BF_{10} = 1.77E23$, but the effect was somewhat smaller than the answer-before condition.

Table 5 The results of the lens model by answer condition in Experiment 4

	Answer before	Answer after
Resolution (Pearson correlation)	.48 (.20)	.47 (.21)
Validity	.34 (.13)	.35 (.12)
Utilization	.55 (.16)	.48 (.16)
G (optimal weighting)	.83 (.29)	.84 (.31)

We compared the resolution of judgments of others' knowledge across conditions, which are shown in Table 5. A two-sample t test showed that the resolution between the answer-before condition did not differ from the answer-after condition, $t(349) = 0.47$, $p = .64$, $d = .05$, $BF_{10} = 0.13$. The Bayes factor indicates strong evidence in favor of the null hypothesis that the two samples are equivalent.

Utilization and validity of cues As in the first two experiments, we computed the utilization and validity of one's metacognitive cues for predicting others' knowledge using the lens model of metacognition (Bröder & Undorf, 2019). The means are displayed in Table 5. A 2 (measure: utilization vs. validity) × 2 (condition: answer before or answer after) showed a significant interaction, $F(1, 349) = 20.92$, $p < .001$, $\eta_p^2 = .057$, $BF_{10} = 3{,}073.45$, a significant effect of measure, $F(1, 349) = 333.43$, $p < .001$, $\eta_p^2 = .49$, $BF_{10} = 1.01E50$, and a significant effect of condition, $F(1, 349) = 6.69$, $p = .01$, $\eta_p^2 = .019$, $BF_{10} = 1.34$. The validity of cues did not differ between condition, $t(349) = 0.84$, $p = .40$, d

= 0.08, BF_{10} = 0.17, but participants in the answer-before condition utilized their cues more strongly than those in the answer-after condition, $t(349) = 4.30$, $p < .001$, $d = 0.44$, $BF_{10} = 707.44$. G, the degree of optimal weighting between validity and utilization, did not differ between conditions, $t(349) = 0.13$, $p = .90$, $d = 0.03$, $BF_{10} = 0.12$.

Discussion

Experiment 4 directly compared estimates of others' knowledge between a group that was required to answer the question before estimating and one that was not required to explicitly answer the question. We intentionally trained participants across conditions on the trivia answers unevenly so that the validity of their knowledge for predicting novices' knowledge was impaired. Answering conditions changed how much participants utilized their own experiences when estimating others' knowledge. More specifically, participants required to answer before estimating others' knowledge utilized their own experiences to estimate others' knowledge more heavily than those who were not required to do so. This difference replicates the comparisons between Experiments 2 and 3. Further, the differences in utilizing one's own experiences for predicting others' knowledge affected calibration, where those in the answer-before condition displayed lower calibration for unstudied items, but higher calibration for twice and thrice studied items compared with those in the answer-after condition. In other words, overall levels of estimates were more tightly tied to the estimators' knowledge in the answer-before condition than the answer-after condition.

Judgment conditions changed how salient one's own experiences were when estimating others' knowledge. By not requiring estimators to answer the questions first, we reduced the salience of their ability to answer each question. Reducing the salience of their own knowledge diminished their utilization of their own cues when producing estimates about others. Participants in the answer-after condition could have engaged in covert retrieval of the answers before estimating others' knowledge in Experiments 3 and 4; however, the data suggest that participants did not do this consistently, as removing the requirement to answer first reduced the utilization of one's own experiences to predict others' knowledge.

Reducing the utilization of one's own experiences did not improve the resolution of estimates (or the optimal weighting of cues), even under circumstances in which participants were unevenly trained on the trivia questions. Uneven exposures to the trivia answers provides a strong test of whether one's own knowledge causes impairments to social judgments because the uneven training across questions in this experiment intentionally reduced the validity of one's own knowledge for predicting untrained novices' knowledge. Despite the answer-after condition utilizing their own experiences to predict others' knowledge less than the answer-before condition, answer conditions did not affect the resolution between groups. These data suggest that reducing the utilization of one's experiences to predict others' knowledge does not improve the accuracy of judgments of others' knowledge. Reductions in utilizing cues related to one's own knowledge likely introduces other noise into judgments, especially when estimators do not have strong theories about how to shift their estimates. In other words, as learners shift away from utilizing their own cues, they do not have diagnostic cues about others' knowledge to replace their own cues. The lack of valid cues about others' knowledge seems to be a significant impairment to accuracy, rather than overutilization of one's own experiences.

General discussion

We tested how and why the "curse of knowledge" impacts judgments of others' knowledge across four experiments. Estimates of others' knowledge became less accurate as estimators gained knowledge of the trivia answers when estimators were required to answer the questions first. More specifically, estimators produced greater overestimates (i.e., worse calibration) of novices' knowledge with additional learning, and estimators' ability to judge which questions were easy and which were difficult (i.e., resolution) became less accurate with additional learning. However, learning the trivia answers did not inevitably change the accuracy of judgments of others' knowledge. The impact of one's own knowledge on the accuracy of estimates of others' knowledge was reduced when the salience of their own experiences was diminished.

Learning across rounds impacted calibration and resolution of estimates of others' knowledge differently. In Experiment 2, calibration worsened across rounds only when learners received feedback, while resolution worsened across rounds regardless of the presence of feedback. The diverging results suggests that different cognitive processes underlie calibration and resolution of metacognitive judgments about others, just as different cognitive processes underlie calibration and resolution of metacognitive judgments about oneself (for a more thorough review of cues that affect calibration and resolution differently, see Rhodes, 2016). Average metacognitive judgments for self and others may both be anchored near the midpoint of the response scale (e.g., Connor et al., 1997). With additional learning, those judgments may slightly increase (causing worse calibration in the current experiments) due to increased fluency of answer retrieval (Birch et al., 2017) or anchoring too heavily in one's ability to answer (Nickerson et al., 1987). Our current results cannot cleanly distinguish between competing explanations of impaired calibration across learning. Yet our

data do suggest that reducing the salience of one's own experiences (by not requiring participants to answer questions before estimating others) can reduce biases in calibration caused by personal knowledge.

While our data cannot disentangle competing explanations for biases in calibration, our results can distinguish between the mechanisms underlying impaired resolution across rounds. The results from these four experiments suggest which mechanisms underlie impairments to resolution with increased exposures to the trivia questions, which is still debated. Inhibition (or anchoring-and-adjustment) explanations suggest that people have difficulty fully suppressing or inhibiting their own knowledge when estimating a novice's perspective (Bayen et al., 2007; Epley & Gilovich, 2001; Groß & Bayen, 2015; Horton & Keysar, 1996; Keysar, et al., 2000; Nickerson, 2001). In anchoring-and-adjustment models of perspective taking, people initially anchor in their own perspective and subsequently adjust away from it. Anchoring is egocentric and automatic, but adjustment requires effortful inhibition and time-consuming perspective monitoring (Horton & Keysar, 1996; Keysar et al., 2000; Keysar et al., 2003). Anchoring-and-adjustment theories of perspective taking suggest that biases in estimates of others' mental states primarily arise because people have difficulty inhibiting their own knowledge and fail to adequately adjust away from their own perspectives (Bayen et al., 2007; Lagattuta et al., 2010; Lagattuta et al., 2014). Research across many communication tasks is largely consistent with this model (e.g., Brown-Schmidt, 2009; Keysar et al., 2000; Lin et al., 2010; Ryskin et al., 2015). However, the data shown across our four experiments somewhat contradict the inhibition explanation for impairments in resolution during knowledge estimation. If failure of inhibition underlies resolution impairments, estimates should get progressively less accurate across rounds *because* people do not reduce their utilization of their own experiences enough across learning. Evidence from the current experiments, however, suggests that estimators reduced their utilization of their own experiences adequately across learning. In fact, estimators reduced utilization of their own experiences to a greater degree than the validity of those experiences dropped. Further, experimentally reducing utilization of one's own knowledge in Experiment 4 did not improve the resolution of participants' judgments.

An alternative (and somewhat related) mechanism underlying the impairments in resolution during perspective taking is fluency misattribution, such that the fluency with which knowledge comes to mind is misattributed to the information being easier than it is (Birch et al., 2017; Harley et al., 2004). In this mechanism, resolution of estimates of a novice's knowledge may get worse across rounds because participants interpret fluency with the questions and answers to indicate that others know the answers. Errors in estimates of others arise because estimators utilize and interpret fluency with answers inappropriately. In other words, estimators are not failing to inhibit influences of fluency, but are misattributing the implications of that fluency. Our data suggest that fluency with answering trivia questions is not driving impairments in the resolution of judgments of others' knowledge. While the validity one's own cues decreases with learning, estimators correspondingly decrease their utilization of their own fluency answering the questions.

Finally, cue-utilization theories of perspective taking suggest that a lack of cues about others' knowledge or the misuse of cues about others' knowledge causes inaccuracies in the resolution of estimates of others' knowledge. The cue-utilization theory of perspective taking posits that estimating others' mental states is an inferential process (Kelley & Jacoby, 1996; Thomas & Jacoby, 2013; Tullis, 2018) in which people deduce social judgments from a variety of diagnostic, available, and salient cues (Bem, 1972; Koriat, 1997; Nelson et al., 1998). The utilization (and underlying availability) of valid cues changes the accuracy of judgments about others. For example, feedback about the correct answer can boost accuracy of resolution because the correct answer can provide additional valid cues about the difficulty of a question, including answer familiarity and question-to-answer associative strength, that boost the relation between normative difficulty and predictions. One's own ability to answer a question and the fluency with which one retrieves answers may also serve as salient cues about what others know (Birch et al., 2017; Harley et al., 2004). Relying one one's own knowledge to estimate others' knowledge is a quick and easy heuristic which can generally be adaptive (Gigerenzer & Gaissmaier, 2011; Nickerson, 2001). One's egocentric experiences (e.g., own knowledge and own test time) can often be a good proxy for those others' mental states (Dawes, 1989; Hoch, 1987). The cue-utilization perspective suggests that estimators' own experiences with the trivia questions accurately reflect their peers' knowledge during their first exposure to the trivia questions; however, as they gain exposure and learn the answers, their personal experiences become less valid predictors of novices' knowledge. In other words, experience with the trivia questions reduces the validity of one's own experiences for predicting novices' knowledge. Estimators correspondingly reduced their utilization of their ability to answer each question and the time need to answer as they learned the answers. These results replicate prior research showing that people shift which metacognitive cues they utilize when predicting others' knowledge if they recognize that their own experiences do not reflect others' experiences (e.g., Ames, 2004; Krueger, 1998). Estimators may be able to reduce the reliance on their own knowledge in our experiments because they may be able to easily attribute their knowledge to the specific study trials; reducing reliance on one's special or unique

knowledge may be more difficult in the real world because estimators would have to understand what aspects of their knowledge are shared and what aspects are unique. If we were to delay the estimation task so that estimators forget that they have learned the information within the task, their estimates of others may be more heavily biased toward their newly acquired knowledge.

Further, our results show that, as estimators reduce the utilization of their own experiences to predict others' knowledge, noise in judgments consistently increases (G decreases). In other words, the optimal weighting of available metacognitive cues weakens with learning. So, not only does the validity of cues drop across rounds, but estimators fail to appropriately weigh the remaining valid cues with learning. Reducing utilization of one's own experiences when making estimates allows other, nondiagnostic cues (or noise) to influence judgments. When taking perspective of others more broadly, people often lack direct information about others' mental states (Tullis & Fraundorf, 2017) and must infer those states from whatever information is available and salient. The lack of valid cues about others' knowledge, rather than utilizing one's own perspective too heavily, can introduce noise into judgments and impair resolution across rounds. In fact, when estimators receive valid cues about specific people's experiences, the accuracy of predictions about others' mental states significantly improves (Jameson et al., 1993; Vesonder & Voss, 1985). The presence (or absence) of valid cues about others' knowledge likely dictates the accuracy of metacognitive judgments. For example, in real world classrooms, the ability of teachers to accurately judge their students' knowledge depends on which cues are available to those teachers (Oudman et al., 2018).

As estimators reduce the utilization of their experiences to predict others' knowledge, they may increase their utilization of abstract theories about what makes trivia questions difficult. When estimators discount their own experiences (and reduce utilization of them), they may incorporate theory-based judgments about the difficulty of questions for others (Kelley & Jacoby, 1996; Thomas & Jacoby, 2013). Theory-based judgments may involve a deliberate analysis of the trivia questions, including beliefs about the objective qualities of the question itself (e.g., Koriat & Bjork, 2006; Koriat et al., 2006). The impact of shifting from experiential cues (like one's ability to answer and test time) to theory-based cues on accuracy of judgments is determined by the validity of those theories. Prior research and the current results suggest that theory-based cues may be less valid for predicting others' knowledge than one's own experiences (Kelley & Jacoby, 1996).

Accurately predicting the normative difficulty of questions can be a vital skill. Estimating others' knowledge may be particularly important for teachers, who utilize estimates about students' knowledge to adapt their instruction to their students' needs (Alvidrez & Weinstein, 1999; Shavelson, 1978). To make effective instructional decisions, teachers' judgments of students' knowledge need to be accurate (Klug et al., 2013). For example, if a teacher can predict the normative difficulty of topics for novices in their class, they may be better able to structure the activities to support student learning (see Sadler et al., 2013). The ability of teachers to estimate real-world student knowledge is likely more complex and may include greater theory-based cues about difficulty than are present within our artificial lab contexts. In fact, a wide variety of metacognitive cues affect the accuracy of teachers' judgments of student learning, including professional development (Thiede et al., 2015), knowledge of the students (Oudman et al., 2018), and instances of formative assessments (Thiede et al., 2018; for a metanalysis of factors that affect the accuracy of teachers' estimates, see Südkamp et al., 2012). Just as accurate monitoring of one's own learning helps students make effective study choices and ultimately supports one's own learning (Thiede et al., 2003; Tullis & Benjamin, 2011), accurate predictions about others' knowledge may help support teaching, communication, and persuasion of others. Our experiments show that increased knowledge can impair both the calibration and resolution of judgments of others' knowledge because cues related to one's own experiences become less predictive of novices' knowledge throughout learning. Estimators reduce the utilization of their own experiences and introduce greater noise into their judgments about others. Inaccuracies in judgments about others and the "curse of knowledge" may result from a lack of valid cues about others, rather than tying estimates too strongly to one's own experiences or misinterpreting fluency with answers. Ultimately, the cue-utilization framework for predicting others' knowledge may suggest methods and conditions to promote accurate utilization of cues to effectively monitor and control others' learning.

Appendix A

In the main text, we present the results of ANOVAs to test the central hypotheses. Here, we present the results of a linear models to test those same hypotheses. Linear models account for the ordered sequence of rounds (rather than the nominal categories of rounds in an ANOVA) and typically have more power than ANOVAs. A significant constraint with these particular models is that they assume *linear* changes across repetitions. In other words, linear models predict an equal change between Rounds 1 and 2 as between Rounds 2 and 3. This assumption may not accurately reflect changes across rounds, especially given that participants'

ability to answer questions does not appear to be linear. Given these particular weaknesses of this analytic approach, we report the results of the linear regressions for the central analyses to account for the ordered nature of rounds. We used the "lmer" package for each analysis, with participant as a random effect. The analyses presented here largely replicate the results provided in the main results section.

Experiment 1

Metacognitive accuracy. We first examined the impact of rounds on measures of metacognitive accuracy. Round was coded as 1, 2, or 3. As shown in Table 6, calibration worsened across rounds, and as shown in Table 7, resolution significantly decreased across rounds.

Utilization and validity of cues. We next examined whether the utilization and validity of cues changed across rounds. In the model shown in Table 8, round was coded as 1, 2, or 3, and measure was coded as 1 for validity and 0 for utilization. The model indicates that both utilization and validity decreased across rounds, but that validity decreased at a slower rate than utilization.

Finally, the fixed effects model predicting G (optimal weighting of cues) by round shows a significant impairment across rounds, as shown in Table 9.

Discussion. The results of the linear models replicate those reported in the main text. First, calibration and resolution became worse across rounds. Second, the validity of one's own experiential cues dropped across learning, but utilization of those same cues dropped more significantly. Finally, G decreased across rounds, indicating greater noise in weighing one's cues for estimating others' knowledge.

Table 6 Fixed effect estimates for mixed effects model of calibration (proportion estimate minus proportion normative difficulty) in Experiment 1 ($N = 393$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.069	0.015	4.69	<.001
Round	0.018	0.004	4.06	<.001

SE = standard error

Table 7 Fixed effect estimates for mixed effects model of resolution (Pearson correlation) in Experiment 1 ($N = 393$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.571	0.018	31.79	<.001
Round	−0.028	0.007	4.30	<.001

SE = standard error

Table 8 Fixed effect estimates for mixed effects model of validity and utilization in Experiment 1 ($N = 786$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.712	0.020	35.82	<.001
Round	−0.138	0.009	15.90	<.001
Measure	−0.133	0.027	5.00	<.001
Round × Measure	0.042	0.012	3.44	<.001

SE = standard error

Table 9 Fixed effect estimates for mixed effects model of G in Experiment 1 ($N = 393$).

Fixed effect	B	SE	t	p
Intercept (baseline rating)	1.10	0.052	21.19	<.001
Round	−0.145	0.024	6.13	<.001

SE = standard error

Experiment 2

Metacognitive accuracy. We first examined how round and feedback affected calibration and resolution. The condition with feedback is coded with a 1, while that without is coded as a 0. As shown in Table 10, calibration was significantly worse in the feedback condition. Further, the interaction of round by condition shows that calibration became significantly worse across rounds in the feedback condition, but did not change in the no-feedback condition.

Next, we examined how resolution (the Pearson correlation between estimates and normative difficulty) changed with round and feedback condition. The results of the linear model are shown in Table 11 and show that feedback improved resolution (compared to the no-feedback condition). Further, resolution decreased with round across each condition.

Utilization and validity of cues. As in Experiment 1, we next examined how the rounds affect the validity and utilization of cues. We computed the models separately for the feedback and no-feedback conditions, so that we can more clearly interpret the results. In the model shown in Table 12,

Table 10 Fixed effect estimates for mixed effects model of calibration in Experiment 2 ($N = 792$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	−0.00206	0.01562	0.132	.90
Round	0.00303	0.00523	0.579	.56
Feedback Condition	0.04434	0.01598	2.774	.006
Round × Feedback Condition	0.01484	0.00740	2.007	.045

SE = standard error

Table 11 Fixed effect estimates for mixed effects model of resolution in Experiment 2 ($N = 790$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.465	0.024	19.03	<.001
Round	−0.023	0.010	2.22	.027
Feedback Condition	0.106	0.031	3.41	<.001
Round × Feedback Condition	−0.002	0.014	0.12	.90

SE = standard error

Table 12 Fixed effect estimates for mixed effects model comparing validity versus utilization in Experiment 2 ($N = 792$)

Fixed effect	*B*	*SE*	*t*	*p*
Feedback condition				
Intercept (baseline rating)	0.725	0.023	31.78	<.001
Round	−0.125	0.010	12.59	<.001
Measure	−0.099	0.010	3.26	.001
Round × Measure	0.021	0.014	1.50	.14
No-feedback condition				
Intercept (baseline rating)	0.582	0.019	30.08	<.001
Round	−0.012	0.008	1.59	.11
Measure	−0.092	0.023	3.89	<.001
Round × Measure	0.015	0.011	1.37	.17

SE = standard error

round was coded as 1, 2, or 3, while measure was coded as 1 for validity and 0 for utilization. In the feedback condition, participants utilized the cues more heavily than the validity and both utilization and validity dropped to the same degree across rounds. For the no-feedback condition, participants utilized cues more heavily than their validity, but round did not impact these weights.

Finally, we examined how the optimal weighing of available metacognitive cues changed across rounds. As shown in Table 13, G decreased across round for both the feedback and no-feedback conditions.

Table 13 Fixed effect estimates for mixed effects model for G in Experiment 2 ($N = 393$)

Fixed effect	*B*	*SE*	*t*	*p*
Feedback				
Intercept (baseline rating)	1.02	0.061	16.67	<.001
Round	−0.16	0.03	5.27	<.001
No-feedback				
Intercept (baseline rating)	0.926	0.045	20.68	<.001
Round	−0.06	0.020	2.82	.005

SE = standard error

Discussion. The results of the linear models largely replicate those presented in the main text. First, calibration became worse across rounds for the feedback condition, but did not change in the no-feedback condition. Second, resolution became worse across rounds in both conditions. Next, the utilization and validity of one's own cues dropped across rounds in the feedback condition, but did not significantly change in the no-feedback condition. Feedback changes how valid one's own experiences are for predicting others' knowledge, and estimators appropriately reduced how much they utilize these experiences. When the validity of one's cues do not drop across rounds, estimators do not reduce their utilization of those cues. In other words, reductions in metacognitive accuracy across rounds are not driven by increased overutilization of one's own experiences. Finally, as reported in the main text, the optimal weighting of cues (G) does significant decrease across rounds, yielding greater noise in estimates and ultimately impairing resolution.

Experiment 3

Metacognitive accuracy. In Experiment 3, we examined how round and condition (answer present vs. answer absent) affected calibration and resolution. The answer-present condition is coded as 1, while the answer-absent condition is coded as 0. The results shown in Table 14 indicate that round had no significant impact on calibration. The results indicate a significant main effect of answer condition, as calibration was significantly higher in the answer-present condition than the answer-absent condition.

Finally, we computed the impact of round and feedback presence on resolution. The results, shown in Table 15,

Table 14 Fixed effect estimates for mixed effects model of calibration in Experiment 3 ($N = 786$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.0274	0.0152	1.81	.07
Round	0.0048	0.0046	1.04	.30
Answer Condition	0.0588	0.0140	4.19	<.001
Round × Answer Condition	0.0050	0.0065	0.77	.44

SE = standard error

Table 15 Fixed effect estimates for mixed effects model of resolution in Experiment 3 ($N = 785$)

Fixed effect	*B*	*SE*	*t*	*p*
Intercept (baseline rating)	0.428	0.026	16.41	<.001
Round	−0.005	0.010	0.51	.61
Answer Condition	0.106	0.031	3.36	<.001
Round × Answer Condition	−0.003	0.015	0.23	.82

SE = standard error

reveal that round had no impact on resolution, but the presence of the feedback significantly improved resolution.

Discussion. Again, the results of the linear models presented here replicate those presented in the main text. Neither calibration nor resolution was affected by round when participants were not required to answer the questions before estimating others' knowledge. Further, when the answer was present, estimators overpredicted others' knowledge but showed better resolution of their estimates.

Declarations

Competing interests The authors have no competing interests to declare.

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*(4), 731–746.

Ames, D. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*, 340–353.

Bayen, U. J., Pohl, R. F., Erdfelder, E., & Auer, T. (2007). Hindsight bias across the lifespan. *Social Cognition, 25*, 83–97.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6). Academic Press.

Berg, T., & Brouwer, W. (1991). Teacher awareness of student alternate conceptions about rotational motion and gravity. *Journal of Research in Science Teaching, 28*, 3–18.

Birch, S. A. J. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Current Directions in Psychological Science, 14*, 25–29.

Birch, S. A. J., & Bloom, P. (2003). Children are cursed: An asymmetric bias in metal-state attribution. *Psychological Science, 14*, 283–286.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382–386.

Birch, S., Brosseau-Liard, P., Bernstein, D., Haddock, T., & Ghrear, S. (2017). A curse of knowledge in the absence of knowledge? People misattribute their feelings of familiarity when judging how common knowledge is among their peers. *Cognition, 166*, 447–458.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica, 197*, 153–165.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review, 16*, 893–900.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue & Discourse, 2*(1), 11–33.

Brunswick, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Camerer, C. F., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*, 1232–1254.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50–71.

Damen, D., van der Wijst, P., van Amelsvoort, M., & Krahmer, E. (2020). Can the curse of knowing be lifted? The influence of explicit perspective-focus instructions on readers' perspective taking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(8), 1407–1423. https://doi.org/10.1037/xlm0000830

Dawes, R. M. (1989). Statistical criteria for a truly false consensus effect. *Journal of Experimental Social Psychology, 25*, 1–1.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*, 391–396.

Epley, N., & Waytz, A. (2009). Mind perception. In S. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (5th ed., pp. 498–541). https://doi.org/10.1002/9780470561119.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fischhoff, B. (1975). Hindsight = foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Fraundorf, S. H., Diaz, M. I., Finley, J. R., Lewis, M. L., Tooley, K. M., Isaacs, A. M., ... Brehm, L. (2014). CogToolbox for MATLAB [computer software]. *Available from* http://www.scottfraundorf.com/cogtoolbox.html.

Friedrichson, P. J., Abell, S. K., Pareja, E. M., Brown, P. L., Lankford, D. M., & Volkmann, M. J. (2009). Does teaching experience matter? Examining biology teachers' prior knowledge for teaching in an alternative certification program. *Journal of Research in Science Teaching, 46*, 357–383.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378–391.

Ghrear, S. E., Birch, S. A. J., & Bernstein, D. M. (2016). Outcome knowledge and false belief. *Frontiers in Psychology, 7*, 1–6.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*(1), 451–482.

Groß, J., & Bayen, U. J. (2015). Hindsight bias in younger and older adults: The role of access control. *Aging Neuropsychology and Cognition, 22*(2), 183–200.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. Bjork (Eds.), *Handbook of memory and metacognition* (pp. 411–455). Erlbaum.

Halim, L., & Meerah, S. M. M. (2002). Science trainee teachers' pedagogical content knowledge and its influence on physics teaching. *Research in Science & Technological Education, 20*, 215–225.

Hargis, M. B., & Castel, A. D. (2019). Knowing what others know: Younger and older adults' perspective-taking and memory for medication information. *Journal of Applied Research in Memory and Cognition, 8*, 481–493.

Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The "saw-it-all-along" effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(5), 960–968. https://doi.org/10.1037/0278-7393.30.5.960

Heine, S. J., & Lehman, D. R. (1996). Hindsight bias: A cross-cultural analysis. *The Japanese Journal of Experimental Social Psychology, 35*, 317–323.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied, 5*, 205–221.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy. *Journal of Personality and Social Psychology, 53*, 221–234.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Jameson, A., Nelson, T. O., Leonesio, R. J., & Narens, L. (1993). The feeling of another person's knowing. *Journal of Memory and Language, 32*, 320–335.

Kelley, C. M. (1999). Subjective experience as a basis of "objective" judgments: Effects of past experience on judgments of difficulty. In D. Gopher & A. Koriat (Eds.), *Attention and performance* (Vol. 17, pp. 515–536). MIT Press.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic basis for judgment. *Journal of Memory and Language, 35*, 157–175. https://doi.org/10.1006/jmla.1996.0009

Kennedy, J. (1995). Debiasing the curse of knowledge in audit judgment. *The Accounting Review, 70*, 249–273.

Keysar, B. (1994). The illusory transparency of intention: Perspective taking in text. *Cognitive Psychology, 26*, 165–208.

Keysar, B., & Barr, D. J. (2002). Self-anchoring in conversation: Why language users do not do what they 'should.' In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 150–166). Cambridge University Press.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38.

Keysar, B., & Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language, 34*, 89–109.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46. https://doi.org/10.1016/j.tate.2012.10.004

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622. https://doi.org/10.1037/0278-7393.32.3.609

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.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945–959.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643–656.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69.

Krueger, J. (1998). Enhancement bias in description of self and others. *Personality and Social Psychology Bulletin, 24*, 505–516.

Lagattuta, K. H., Sayfan, L., & Blattman, A. J. (2010). Forgetting common ground: Six to seven-year-olds have an overinterpretive theory of mind. *Developmental Psychology, 46*, 1417–1432.

Lagattuta, K. H., Sayfan, L., & Harvey, C. (2014). Beliefs about thought probability: Evidence for persistent errors in mindreading and links to executive control. *Child Development, 85*(659), 674.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551–556.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability in free and cued recall. *Bulletin of the Psychonomic Society, 22*, 497–500.

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

Nelson, T. O., Kruglanski, A. W., & Jost, J. T. (1998). Knowing thyself and others: Progress in metacognitive social psychology. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 69–89). SAGE.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Memory and Language, 19*, 338–368.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science, 10*(5), 168–172.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica, 64*, 245–259.

Oudman, S., Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology, 49*(4), 270–282. https://doi.org/10.1026/1618-3169.49.4.270

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes, 67*, 49–58.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 90–117). Oxford University Press.

Ryskin, R. A., Benjamin, A. S., Tullis, J. G., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General, 144*, 898–915.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal, 50*, 1020–1049.

Shavelson, R. J. (1978). Teachers' estimates of students' states of mind and behavior. *Journal of Teacher Education, 29*(5), 37–40.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition, 41*, 1000–1011.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods, 45*, 1115–1143.

Thiede, K. W., Anderson, C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the

accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76*, 106–115. https://doi.org/10.1016/j.tate.2018.08.004

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., Oswalt, S., ... Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*, 290–302.

Thomas, R. C., & Jacoby, L. L. (2013). Diminishing adult egocentrism when estimating what others know. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 473–486.

Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & Cognition, 46*, 1360–1375.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*, 109–118.

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language, 95*, 124–137.

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24*(3), 363–376.

Wieman, C. (2007). New formula for science education. *Physics World, 20*, 10.

Open practices statement

Data from the experiments are available on the Open Science Framework (https://osf.io/2ngbq/?view_only=ada6614377a24bc797df3046dcee2872).

None of the experiments was preregistered.