Mem Cogn
DOI 10.3758/s13421-011-0130-z

The costs and benefits of memory conformity

Antonio Jaeger · Paula Lauris · Diana Selmeczy · Ian G. Dobbins

Abstract We examined the influence of external recommendations on memory attributions. In two experiments, participants were led to believe that they were viewing the responses of two prior students to the same memoranda they were currently judging. However, they were not informed of the reliability of these fictive sources of cues or provided with performance feedback as testing proceeded. Experiment 1 demonstrated improvement in the presence of reliable source cues (75% valid), as compared to uncued recognition, whereas performance was unaltered in the presence of random cues provided by an unreliable source (50% valid). Critically, participants did not ignore the unreliable source, but instead appeared to restrict cue use from both sources to low-confidence trials on which internal evidence was highly unreliable. Experiment 2 demonstrated that participants continued to treat an unreliable source as potentially informative even when it was predominantly incorrect (25% valid), highlighting severe limitations in the ability to adequately discount unreliable or deceptive sources of memory cues. Thus, under anonymous source conditions, observers appear to use a low-confidence outsourcing strategy, wherein they restrict reliance on external cues to situations of low confidence.

Keywords Memory · Memory conformity · Cueing · Recognition

In the laboratory, great efforts are taken to ensure that decision strategies do not inflate estimates of recognition memory accuracy. For example, in a typical recognition memory test, studied and nonstudied items are randomly intermixed and equiprobable. While such procedures are necessary for understanding baseline recognition abilities, they are arguably quite artificial, because recognition judgments outside the laboratory should make use of a host of situational cues in order to improve performance. For example, when attempting to identify someone at a high school reunion as a former classmate (as opposed to, say, the spouse of a classmate), one might use cues such as the general context of the event (are most attendees familiar?) or the person's age (is it consistent with my peer group?), or ask a friend's explicit opinion (does he or she claim to recognize the person?). Such contextual cues are generally predictive and could decrease the likelihood of awkward mistakes.

Here we examine how observers incorporate the judgments of anonymous others into their recognition decisions by providing them with the recognition judgments of two fictitious others during the course of testing. Before describing the optimal solution to this task, we briefly review work in social psychology that examines similar questions under the rubric of "social conformity." Pioneering work by Asch (1955) examined how the judgments of others influenced one's perceptual decisions. In a series of experiments, 7–9 participants were assembled in a room to take part in a fictional visual perception experiment in which they indicated whether vertical lines had the same or different lengths. Of these people, only 1 was a naïve participant, while the rest were confederates who responded almost unanimously according to a prearranged plan. Responses were verbalized, and the naïve participant responded last. Even though participant accuracy when tested in isolation was typically close to 100%, the research demonstrated that when confederates gave misinformation,

A. Jaeger (✉) · P. Lauris · D. Selmeczy · I. G. Dobbins
Department of Psychology, Washington University,
1 Brookings Drive,
St. Louis, MO 63130-4899, USA
e-mail: antonio.jaeger@gmail.com

Published online: 20 July 2011

the naïve participants' scores were clearly impaired by the aberrant group consensus.

A host of studies have examined these social conformity effects on a variety of judgments (e.g., Baron, Vandello, & Brunsman, 1996; Hoffman, Granhag, Kwong See, & Loftus, 2001), and a growing number of studies have examined these effects on explicit recognition judgments (Allan & Gabbert, 2008; Axmacher, Gossen, Elger, & Fell, 2010; Gabbert, Memon, & Wright, 2007; Meade & Roediger, 2002; Reysen, 2005; Schneider & Watkins, 1996; Walther, Bless, Strack, Rackstraw, Wagner, and Werth 2002; Wright, Gabbert, Memon, & London, 2008; Wright, Self, & Justice, 2000). In general, the findings have shown that participants will shift their recognition memory decisions toward confederates' even when those confederates' reports are incorrect, a phenomenon termed "memory conformity." Because most research on memory conformity considers laboratory findings in the light of eye-witness testimony situations, memory conformity has generally been characterized as undesirable. For example, Walther et al. commented that it was "not enough that our cognitive system is apparently faulty, now it also appears that other people, perhaps people we do not like or do not even know, are able to control our recollections." Thus, the emphasis of the memory conformity studies has typically been on the degree to which observers allow themselves to be negatively impacted by others who are purposefully inaccurate and deceptive. This emphasis has perhaps had the unfortunate effect of obscuring the fact that the use of external cues or information during the course of recognition judgments is the statistically optimal approach under Bayes's theorem and the closely related theory of signal detection (Macmillan & Creelman, 2005).

What is the optimal decision rule when given a recommendation while making a recognition judgment? Ideally, two pieces of information are required, namely (1) the long-term validity of the recommender and (2)internal recognition evidence that either corroborates or refutes that recommendation. Furthermore, the internal evidence needs to reflect the relative likelihoods of the perceived memory experience under the possibility that the item was studied relative to the possibility that it was instead unstudied, a value termed the "likelihood ratio" (Macmillan & Creelman, 2005). Given these two pieces of information, the participant can maximize performance when recommendations are available. As a concrete example, say that the long-term validity of the source is 75%, and on a particular trial the observer perceives a level of familiarity consistent with 1.5 to 1 odds that an item was studied versus new. If the source also recommends an "old" judgment, then, through Bayes's theorem, the posterior or adjusted odds that the item would in fact be old are given by the validity of the source (expressed as prior odds) multiplied by the likelihood ratio evidence, $3/1$ * $1.5/1$ or 4.5 to 1. Given this information, the observer should respond "old," and fairly confidently so. Now consider the case in which the source recommends a "new" judgment. Here, the prior odds of the item being old are 1 in 3, and so the posterior odds are $1/3$*$1.5/1$ or 1.5 to 3. In this situation, even though the observer's internal evidence suggests that the item is old (1.5 to 1), he or she should nonetheless respond "new" because the internal evidence is insufficient to countermand the prior odds provided by the external source.[1] Critically, because the source provides evidence that is diagnostic at the trial level, the observer using Bayesian reasoning will be more accurate in the presence of the cues than in their absence.

Using the mechanics of signal detection theory, one can calculate the expected elevation in accuracy under cues of different validity given an observer's baseline accuracy. For example, an observer with a d' of 1.0 in standard conditions would ideally demonstrate a discriminability of 1.53 in the presence of an external recommender that was known to be 75% valid. Thus, the current decision problem can be envisioned in two different ways under signal detection theory. One way is to assume that the observer uses a fixed optimal decision criterion based on a likelihood ratio evidence axis. Under this approach, the assumed evidence distributions are different for cued and uncued trials, with the former being separated by a greater distance. Evidence distributions under the cued conditions reflect the optimal combination of evidence from all available sources prior to the decision stage, and thus the evidence distributions are farther apart during cued trials, reflecting the increased information provided by the external source (see Glanzer, Hilford, & Kim, 2004, for an analogous approach). Alternatively, under an item strength decision axis model, the evidence distributions remain fixed across the cued and uncued trials, reflecting the assumption that the observer's retrieval evidence does not change as a function of the external recommendations. Under this model, the observer shifts his or her decision criterion on a trial-by-trial basis, placing it in the optimal location given the specific external cue. Although psychologically different, these models are indistinguishable behaviorally, and both predict the same improvement in measured performance under cued versus uncued conditions.

Returning to the present research paradigm, the question is "What should the observer do if the validity of the source of recommendations is entirely unknown?" Here, Bayes's theorem is of little help because the prior (i.e., the validity of the source) is unknown. Put simply, one cannot optimally

[1] We thank Gordon Pitz of University of North Carolina: Chapel Hill for valuable discussion regarding the correspondence between Bayesian reasoning and signal detection theory.

incorporate information from a source that is anonymous. Indeed, under this approach the observer would be assumed to rely solely on his or her internal evidence to guide judgments. This prediction, however, ignores an important source of potential knowledge—namely, the observers' self-knowledge about the relationship between confidence and the likelihood of success (a.k.a. *metamemory*). If an observer could restrict reliance upon external anonymous cues to those trials on which subjective confidence is low, then prior research suggests that it is likely he or she could benefit from even anonymous sources. This is because recognition performance during low-confidence reporting or during subjective reports of guessing tends to be close to chance. For example, in an analysis of 86 experimental conditions from 23 experiments, Gardiner, Ramponi, and Richardson-Klavehn (2002) contrasted the accuracy for subjective reports of remembering, knowing, and guessing. Critically, the proportion of correct responses for subjective experiences of guessing was indistinguishable from chance. Given this, if a participant restricted his or her use of an external source to such subjective experiences, then the only way performance could fail to improve is if the external source was random (viz., 50% cue validity), and the only way performance could actually decline is if the external source was systematically incorrect or deceptive. We call this putative strategy *low-confidence outsourcing*.

To investigate whether observers adopt this strategy, we expanded a procedure developed by O'Connor, Han, and Dobbins (2010) that used anticipatory cues during recognition and is conceptually similar to the attentional cueing procedure developed by Posner, Snyder, and Davidson (1980). During the explicit mnemonic cueing procedure, each recognition memory probe is preceded by a single probabilistic cue that forecasts its likely status (e.g., "likely old" or "likely new"). In the version used here, we provided the participants with fictional old/new classification reports ("old" or "new") of participants who ostensibly completed the same recognition tests, and we *omitted* the feedback that was presented in O'Connor et al. (2010), to discourage participants from learning about source validity.

In two experiments, we examined several questions. First, we examined whether in the absence of feedback and knowledge of source reliability, observers are able to use external sources to improve their performance during recognition. Based on the low-confidence outsourcing strategy discussed above, we predicted that even without feedback, participants might benefit from a source that was generally valid. The second question we explored was whether observers would naturally distinguish between external sources of differing reliability, preferentially weighting a source that was reliable and ignoring a source that was wholly unreliable/random (Exp. 1) or systematically incorrect/deceptive (Exp. 2). Because participants were not explicitly informed about the relative reliabilities of the sources, nor provided with feedback following responding, we anticipated that it might be relatively difficult to appropriately favor one source over the other. Finally, since in the present paradigm actual confederates were not employed, but merely the fictional reports of others, these experiments indirectly test whether direct social interaction is necessary for robust "conformity" effects during standard recognition tests.

Experiment 1

Here we modified the explicit mnemonic cueing paradigm of O'Connor et al. (2010) to examine whether observers, in the absence of feedback or external indications of source reliability, would naturally distinguish between reliable and unreliable sources of memory cues, and hence use those cues accordingly. To examine this in a controlled fashion, participants were led to believe that they would view the answers of two prior anonymous students who had completed the recognition test that they were currently taking. The fictional students' answers were primarily shown in isolation; however, a secondary manipulation also examined the effects when both students' reports were provided on a given trial. Critically, one of the student sources was wholly unreliable (50% reliability), whereas the other was moderately reliable (75% reliability). Ideally, one might expect the participants to rely moderately on the reliable student and to completely discount the unreliable student; however, this did not occur.

Method

Participants Experiment 1 included 23 Washington University undergraduate students (18–21 years old; 16 females, 7 male) who participated in return for course credit. Two participants were excluded due to chance performance, leaving 21 for analysis. Informed consent was obtained in accordance with the Institutional Review Board of the university.

Materials A total of 480 words were randomly drawn for each participant from a pool of 1,216 total words. From this set, three lists of 160 items (80 old and 80 new items for each cycle) were used in three study–test cycles. The items in the pool had on average 7.09 letters and 2.34 syllables, with a Kučera and Francis (1967) corpus frequency of 8.85.

Procedures Participants were seated at separate computer consoles and tested using standard PCs with a maximum of 4 participants per session. During encoding, participants

indicated whether serially presented words had more than one syllable, with the cue "More than 1 syllable? Yes or No" appearing beneath each word. This task promotes shallow encoding of the items, consequently avoiding ceiling performance in the following recognition memory test. Participants were given 1.5 s. to respond. If they failed to respond within this length of time, the response was scored as incorrect and the next trial began. At test, the 80 studied items were randomly intermixed with 80 new items and presented consecutively for a recognition judgment using a 6-point confidence rating scale (*very confident old*, *somewhat confident old*, *guessing old*, *guessing new*, *somewhat confident new*, and *very confident new*). The key assignment was counterbalanced between participants, and responses were self-paced. During the cued trials, each probe was preceded by cues ("old" or "new") that probabilistically forecasted the study status of the upcoming probe. Participants were told that the cues were the actual responses given by two anonymous students who had taken part on the same test previously. In reality, the anonymous students were fictional, and the cues were programmed to achieve the two levels of validity. The cues preceded the probe by 1 s and were kept on the screen until the probe's offset. They were shown on the top left and top right quadrants of the computer screen underneath identifiers for the two fictional students ("Student A" and "Student B"). Cues given by one of the fictive students were generally valid (75% valid), and we refer to this as the "reliable source." In contrast, cues given by the other student ("unreliable source") were random (i.e., 50% valid). The positions on the screen of the reliable and unreliable sources was counterbalanced across participants.

In the subset of trials on which cues from both students were simultaneously provided, the cues from the unreliable source remained 50% valid, whereas the cues from the reliable source became 100% valid. Trials on which the cues from one student were presented ("single-cue" trials) and on which cues from both students were presented ("double-cue" trials) were intermixed with trials on which cues from neither student were presented ("uncued" trials). Single-cue trials comprised 50%, double-cue trials comprised 37.5%, and uncued trials comprised 12.5% of the total number of trials. The experiment duration remained under 1 h, and participants were allowed to rest between each of the three study–test cycles. Immediately following the final study–test cycle, participants were administered a questionnaire probing their knowledge of the differences between the fictional students' accuracy. The questionnaire consisted of five questions focusing on the observers' awareness of the sources' reliability differences and on whether or not the sources influenced their responses. Following this, participants were debriefed and dismissed.

Results and discussion

Single cue trials To initially examine whether participants were able to improve performance using the cues,[2] we contrasted their accuracy on uncued trials (d') with their accuracy on single-cue trials for the reliable and unreliable sources[3] (see Table 1). As revealed by a one-way ANOVA [$F(2, 40) = 8.31$, $MSE = .084$, $p < .001$], performance was altered by the cueing procedure (uncued, reliable source cues, and unreliable source cues). Pair-wise follow-up contrasts confirmed that performance was enhanced in the reliable source condition relative to the uncued condition [$t(20) = 3.62$, $p < .002$], whereas during the unreliable source condition performance did not differ from the baseline, uncued performance ($t < 1$). These findings clearly demonstrate that cues provided by the reliable source benefited recognition performance, whereas cues from the unreliable source did not incur a cost. Critically, this finding demonstrates that observers do not need either feedback or explicit information about the reliability of an external source in order to benefit from cues provided by that source. At first glance, this might lead one to conclude that the observers simply ignored the unreliable source based on

[2] It may strike readers familiar with signal detection theory as odd to talk about observers boosting their d' as a function of factors that influence a decision criterion. This is because under most experimental circumstances decision criterion positioning and evidence values are independent at the trial level. However, this is not the case in the explicit mnemonic cueing paradigm. To illustrate, an ideal observer with a d' of 1 will have a baseline hit rate of .69 and a false alarm rate of .31. With a 75% valid external cue, this observer should shift the decision criterion 1.10 units to the left of neutral for a "likely old" cue ($C = -1.10$) and 1.10 units to the right for a "likely new" cue. This would result in a (.95, .73) hit and false alarm rate under the "likely old" cue and a (.27, .05) hit and false alarm rate under the "likely new" cue. Both of these pairs of values correspond to a d' of 1, and hence one might expect cued and uncued performance to be identical. However, the correct response rates represent a considerably larger proportion of the total test trials than the incorrect response rates, because the cues are 75% valid. In other words, the gains incurred when using the cues occur far more frequently (75% of the trials) than the costs incurred when using the cues (25% of the trials), precisely because the cue is valid. These different proportions must be weighted appropriately when calculating the expected net gain in performance with the use of the cues. Hence, in the present example, the net hit rate of the observer is actually .75 * (.95) + .25 * (.27) = .78, and the net false alarm rate is .25 * (.73) + .75 * (.05) = .22, a gain of 9% in the hit rate and a reduction of 9% in the false alarm rate compared to the baseline values above. This is what yields the expected performance improvement.

[3] For both Experiments 1 and 2, in order to calculate the signal detection measures of accuracy (d') and response bias (C) for all participants, scores for participants with perfect hit or false alarm rates were corrected by adding .5 to the hit and false alarm frequencies and dividing them by $N + 1$, where N is the number of old or new trials (see Snodgrass & Corwin, 1988). For consistency, the correction was also applied to participants without perfect scores. None of the conclusions were altered by instead simply restricting the analyses to the noncorrected data of participants without perfect scores.

Table 1 Uncued, reliable, and unreliable source collapsed accuracy (d') (standard deviations in parentheses)

	Uncued	Reliable Source	Unreliable Source
Experiment 1	1.19 (0.59)	1.53 (0.57)	1.26 (0.54)
Experiment 2	1.16 (0.49)	1.27 (0.47)	1.01 (0.45)

a clear belief that he or she was not skilled at recognition. However, the analysis below and Table 2 show that this was not the case.

If an observer wholly discounts an external source, then his or her performance will be equivalent regardless of that source's recommendations. This was clearly not the case for the participants when responding in the presence of the unreliable source cues, because the tendency to respond "old" (both correctly and incorrectly) increased approximately 10% across the "new" and "old" recommendations given by this source (see Table 2). To test this formally, we calculated the signal detection theory bias measure C for each type of recommendation ("old" or "new") from each source (reliable or unreliable), and performed a two-way repeated measures ANOVA. The analysis demonstrated a main effect of source [$F(1, 20) = 4.36$, $MSE = .024$, $p < .05$], with responding somewhat more liberal under the unreliable than under the reliable source. There was also a prominent main effect of recommendation, with more liberal responding under the "old" versus "new" recommendations [$F(1, 20) = 14.86$, $MSE = .20$, $p < .001$]. Critically, these factors did not significantly interact, suggesting comparable shifts of criterion under both of the sources [$F(1, 20) = 1.78$, $MSE = .067$, $p = .20$] (Table 2). These data demonstrate that the participants did not simply ignore the unreliable source, and indeed they were sizably influenced despite its complete randomness.

Overall, these results demonstrate that the differential accuracy benefit gained from the reliable versus unreliable sources was not the result of participants attending to the former and ignoring the latter. Instead, they used both to a similar extent, accruing gains under the reliable source with no appreciable costs for using the unreliable source. The lack of costs associated with using a wholly random source can only occur if observers restricted their use of the cues from this source to trials on which performance in the absence of cues (i.e., baseline) would have been near random. That is, the only way one can be clearly influenced by a random source (i.e., demonstrate criterion shifts), yet not show a performance decline relative to baseline, is if the influence is restricted to trials where the internal evidence would have also resulted in chance responding. This conclusion is further supported by an analysis of the observers' low-confidence responses during baseline recognition. During these trials, correct response rates were only 58% for the 18 observers with sufficient data. If observers defaulted to the external sources during analogous trials when cues were present, they would benefit from the reliable source (75% vs. 58% correct) but not appreciably suffer from the unreliable source (50% vs. 58% correct). As mentioned above, the putative "low-confidence outsourcing" strategy accounts for the data pattern simply because the reliable source more often provides the correct answer on trials of low subjective confidence than does the unreliable source.

To test this hypothesis further, we used the baseline performance data to predict the cued performance accuracy pattern. Figure 1a shows the percentages of correct responding for each level of expressed confidence during baseline, uncued recognition. As expected, accuracy increases with expressed confidence, and as noted above, low-confidence responding is not much above chance, at 58%. Critically, we can use these baseline confidence percentages to predict the patterns of performance under the two cue sources.[4] Panel b shows the overall percentage correct during baseline, simply collapsing across the confidence data shown in panel a. The remaining two boxplots show the predicted percentages on cued trials if participants simply deferred to the external cues when experiencing low subjective confidence. For example, a participant whose percentage correct during low-confidence baseline trials was 55% would have this value replaced with 50% correct during the unreliable source cue trials and 75% during the reliable source cue trials. These percentages are then appropriately weighted by the relative proportions of trials given low confidence, and the total predicted percentage correct is recalculated, all based on the baseline data. These predicted percentages yield a pattern highly similar to the empirical findings under the cued trials—namely, only a slight decline in performance under the random cues (~1%), but a more prominent increase in performance under the reliable cues (~4%). Critically, these predictions do not rely on differential use of the cues, as participants always default to the external cue when experiencing low confidence. Far from reflecting an undesirable approach, however, the present analysis suggests a very useful strategy capable of operating even when external sources are wholly anonymous and feedback-based learning about source reliability is impossible. Since

[4] It should be noted that one cannot use the confidence data acquired during cued trials to examine whether observers selectively deferred to the external sources during low confidence. This is because the confidence reports during cueing reflect some unknown mix of internal evidence assessment and the influence of the external cue that was provided. For example, confidence could reflect factors such as initial agreement with the external cue. This precludes the use of these data for predicting patterns under the proposed "low-confidence outsourcing" strategy.

Table 2 Experiment 1 and 2 mean "old" response proportions for targets and lures and mean response bias (*C*) according to source cueing (standard deviations in parentheses)

			Reliable Source			Unreliable Source		
		Uncued	"Old"	"New"	Collapsed	"Old"	"New"	Collapsed
Exp. 1	Targets	.74 (.12)	.80 (.14)	.71 (.22)	.78 (.13)	.78 (.12)	.68 (.13)	.73 (.11)
	Lures	.31 (.16)	.39 (.23)	.20 (.12)	.25 (.12)	.32 (.17)	.23 (.09)	.28 (.11)
	C	–.08 (.31)	–.33 (47)	.12 (.44)	.07 (.32)	–.18 (.36)	.12 (.24)	–.02 (.24)
Exp. 2	Target	.70 (.13)	.78 (.11)	.64 (.16)	.74 (.10)	.73 (.13)	.67 (.13)	.68 (.11)
	Lures	.28 (.13)	.41 (.21)	.24 (.14)	.28 (.13)	.32 (.13)	.27 (.14)	.31 (.12)
	C	.04 (.34)	–.27 (.41)	.19 (.36)	–.03 (.28)	–.06 (.28)	.08 (.33)	.01 (.25)

Exp = experiment; "Old" = cue was "old"; "New" = cue was "new"; Collapsed = responses for each cue were collapsed according to source

encountering a source with chance recognition accuracy would be extremely rare, and there is little cost to relying upon such a source during low subjective confidence, the present data demonstrate that "conformity" is beneficial in the long run.

Double-cue trials The above analysis demonstrates that participants actively use the recommendations of a random source, and it also suggests that they do so to a similar extent as with a reliable source. Below we examine whether this pattern continued even when the unreliable source's recommendations occurred alongside the reliable source's recommendations. In short, we examine whether the reliable source was able to mute the influence of the unreliable source.

It is important to note that the design did not combine all possible pairings during the double-cue trials. On all double-cue trials, the reliable source cue was correct (i.e., 100% valid), whereas the unreliable source cues remained random (i.e., 50% valid). Critically, this enables another test of the hypothesis that the observers were largely insensitive to the differential reliability of the two sources, because if this were not the case, the presence of cues from the reliable source should severely mute the influence of the unreliable source. To assess this hypothesis, we calculated the bias of the observers as a function of the recommendations of the unreliable source ("old" vs. "new"). Again, if the presence of the reliable source trumped the recommendations of the unreliable source, bias should then not shift as a function of the unreliable source's recommendations (i.e., it should be ignored). This did not occur. Participants were more liberal on double-cue trials when the unreliable source recommended an "old" response ($C = -.17$) versus a "new" response ($C = .03$) [$t(20) = 3.18$, $p < .01$]. Indeed, the magnitude of

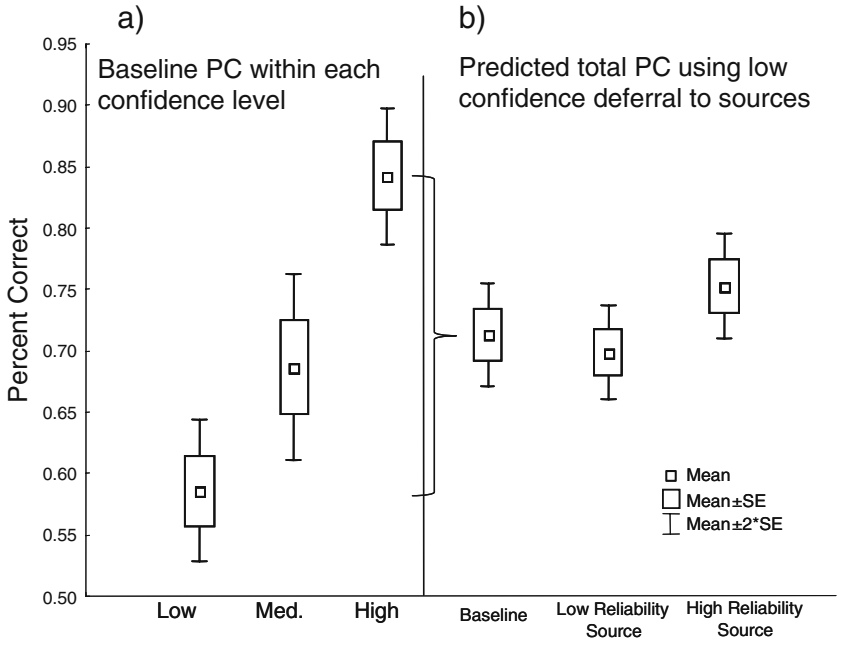

Fig. 1 Expected performance if cues are used in low-confidence trials. (**a**) Percentages of correct responses for each confidence level on uncued trials. (**b**) Raw data (confidence collapsed) from uncued trials (left column) and expected percentages if participants exchanged their low-confidence percentages for complete reliance on the cues from the reliable and unreliable sources (remaining columns). PC, percent correct; Med., medium

this shift was comparable to that seen when the unreliable source was encountered in isolation, and further demonstrates that the observers were largely insensitive to the different utilities of these two sources. Put another way, the data demonstrate that observers attended to whether the unreliable source agreed or disagreed with the reliable source, despite the fact that the former was always correct and the latter random, on these double-cue trials.

Overall, the findings of Experiment 1 yielded two main conclusions. First, observers can capitalize on external cues during recognition, even when the reliability of the external sources is completely unknown and there is no feedback from the environment. They appear to achieve this benefit by using a low-confidence outsourcing strategy whereby they defer to an external source selectively during low-confidence recognition. Because low confidence is associated with near-chance responding in the absence of cues, there is little reason not to use an external source, barring purposeful deception or amnesia on the part of the source.

Despite the utility of the above low-confidence outsourcing strategy, the data also clearly highlight a surprising limitation on the part of the participants. Namely, they appear largely unable to differentiate a reliable from a wholly random source of recognition cues. This was clearly demonstrated by the similar criterion shifts seen in response to "old" and "new" cues provided by the two sources in isolation. While this may seem surprising at first, it is important to remember that feedback was not available during testing. Furthermore, in any given local portion of the test list, the relative accuracy of the two sources would be arguably hard to discern, with the reliable source providing on average 7.5 correct recommendations per every 10 trials and the unreliable source 5 correct recommendations per every 10 trials. Given the known limitations of working memory capacity (Unsworth & Engle, 2007) and the intermixing of recommendations from the two sources, it seems extremely unlikely that observers could actively track these different tendencies (particularly since they were also engaged in demanding episodic judgments). From the observer's limited perspective, there would be frequent agreement between his or her internal assessments and the two external sources, and this agreement (or disagreement) is the only metric available for determining relative source reliability in this design. Thus, the failure of the participants to detect that the unreliable source was wholly random and should be completely ignored is quite understandable.

Experiment 2

Although Experiment 1 demonstrated that observers can capitalize on external cues during recognition, it also showed that observers were largely insensitive to differences in the reliability of the two sources. To examine whether there were limits to this insensitivity, we attempted to make the suspect reliability of the unreliable source even more salient by rendering it systematically incorrect. That is, the unreliable source was modified so that it was correct on only 25% of the trials. Indeed, one could characterize this source as deceptive, in the sense that it generally provided an incorrect answer. There were two related general predictions regarding this manipulation. First, we anticipated that participants might show less reliance on this source than on the reliable source, given the presumably increased frequency with which they would tend to disagree with it. Second, we predicted that they would not actually benefit from the unreliable source. The latter prediction may seem odd, because at first one might question how one could benefit from a source that is only 25% valid. However, because the unreliable source is now incorrect as often as the reliable source is correct, it statistically provides just as much information on the status of each test probe. If an observer agreed with the reliable source on every trial, they would be correct on 75% of the trials. If they also disagreed with the unreliable source on every trial, they would again be correct on 75% of the trials. However, in order to use the unreliable source in this manner, it would require counterfactual reasoning or a contrarian strategy. That is, participants would have to adopt a strategy of systematically disagreeing with the unreliable source specifically on those trials on which their internal evidence led to high subjective uncertainty about the items' memory status. We doubted that they would take this additional cognitive step.

Method

Participants Experiment 2 included 35 Washington University undergraduate students (18–28 years old; 23 females, 12 males) who participated in return for course credit or payment. One participant was excluded due to chance performance, leaving 34 for analysis.

Materials A total of 492 words were randomly drawn for each participant from the same pool of 1,216 words employed in Experiment 1. From this set, three lists of 164 items (82 old and 82 new items for each cycle) were used in three study–test cycles.

Procedures Responses given by the reliable student/source were 75% valid, whereas responses given by the unreliable student/source were 25% valid. In the double-cue trials, the reliable student was 100% correct, the unreliable student 25% correct. The overall proportions of trials were 48.8% with a single cue, 39% with a double cue, and 12.2% uncued.

Results and discussion

Single-cue trials As in Experiment 1, we initially examined whether or not participants were able to improve performance using the cues by contrasting accuracy (d') under the uncued, reliable, and unreliable source trials (see Table 1). As revealed by a one-way ANOVA, cueing clearly influenced accuracy across these conditions [$F(2, 66) = 5.01$, $MSE = .119$, $p < .01$]. Paired follow-up $t$tests revealed that cues from the unreliable source led to a significant decrease in accuracy relative to cues from the reliable source [$t(33) = 3.22$, $p < .01$]; however, neither cued condition reliably differed from baseline performance [reliable source vs. uncued, $t(33) = 1.48$, $p = .15$; unreliable source vs. uncued, $t(33) = 1.64$, $p = .11$]. Although neither source induced a significant departure from baseline accuracy in isolation, the decline in performance under the unreliable relative to the reliable source clearly demonstrates continued reliance on these sources for judgments. This is corroborated by the analysis below.

As in Experiment 1, we assessed the degree to which each source influenced the observers by examining measured criterion (C) under "old" and "new" cue trials for each source using a Source (reliable vs. unreliable) x Recommendation ("old" vs. "new") repeated measures ANOVA (Table 2). There was no main effect of source [$F(1, 33) = 2.14$, $MSE = .038$, $p = .15$] on measured criterion; however, there was a robust effect of recommendation [$F(1, 33) = 28.06$, $MSE = .109$, $p < .001$], with more liberal responding following "old" than following "new" recommendations. Unlike in Experiment 1, these factors interacted [$F(1, 33) = 8.85$, $MSE = .096$, $p < .01$], demonstrating that the degrees of influence differed for the two sources. Post hoc comparisons demonstrated that the interaction occurred because observers responded more vigorously to the cues from the reliable versus the unreliable source. They were more liberal following an "old" cue from the reliable versus the unreliable source [−.27 vs. –.06; $t(33) = 2.86$, $p < .01$] and more conservative following a "new" cue from the reliable versus the unreliable source [.19 vs. .08; $t(33) = 2.12$, $p < .05$].

The question then arises of whether the influence of the reliable and unreliable sources on response bias built up over time or was established early in the experimental session. Since feedback for performance was not provided here, to detect that the unreliable source was systematically giving incorrect cues could have required a considerable number of trials. Given that the present task was divided into three study–test cycles, we conducted an ANOVA on response bias (C) with the factors Cue Recommendation ("old" vs. "new"), Source Reliability, and Study–Test Cycle to verify whether the criterion shifts resulting from the cues given by the unreliable source diminished over time. There was no significant interaction between these factors [$F(2, 66) = 1.54$, $MSE = .109$, $p = .22$], suggesting that the differential reliance on the sources was acquired early in the paradigm (see Table 3).

Although the criterion data clearly indicate that observers were differentially using the two sources (unlike in Exp. 1), they nonetheless continue to point to a major shortcoming in the manner in which participants responded to the unreliable source. That is, participants were more liberal following an "old" than following a "new" recommendation from this source [−.06 vs. .08; $t(33) = 2.24$, $p < .05$]. However, this source was in fact antipredictive or deceptive, such that if it recommended "old," the odds were 3 to 1 that the item would in fact be new. Thus, an ideal observer would demonstrate biases that would be reversed in comparison with the source's recommendations (i.e., the signs on the C measures should be reversed). The failure of participants to do so confirms our prediction that they would not take a contrarian stance in response to a systematically incorrect source. Of the 34 participants, only 5 demonstrated signs on their criterion measures that were reversed in comparison to the cue's recommendations, suggesting a contrarian stance. Thus, while participants were somewhat able to diminish their reliance on this source, the vast majority nonetheless treated it as though it provided valid cues. In short, they displayed a confirmatory bias toward a deceptive source.

Double-cue trials As in Experiment 1, the double-cue trials can be used to see if the presence of the reliable source served to mute the influence of the unreliable source. If the observers completely relied on the reliable source, the recommendations of the unreliable source would have no influence on criterion. Thus, as with Experiment 1, we compared the criterion when the unreliable source indicated "old" to the criterion when the unreliable source indicated "new." If the observer ignored this source in the joint presence of the reliable source, these values should be equivalent. This did not occur, and indeed, the influence of this source when giving "old" and "new" recommendations was comparable to when it was presented in isolation [−.17 vs. .02; $t(33) = 3.00$, $p < .01$]. This remains in the wrong direction, given the antipredictive nature of this source, and the results demonstrate that observers continued to use it inappropriately even when it was presented alongside a genuinely reliable source. When we looked for reversed signs on criterion values that might suggest a contrarian stance, only 3 participants demonstrated reversed signs on the measure C. Thus, again, the data suggest that most observers continued to treat the antipredictive source as actually moderately predictive, and they reinforce the idea that there is a confirmatory bias toward the sources in this experimental design.

Table 3 Experiment 2 mean criterion (*C*) for each study–test cycle according to source cueing (standard deviations in parentheses)

		Reliable Source			Unreliable Source		
	Uncued	"Old"	"New"	Collapsed	"Old"	"New"	Collapsed
Cycle 1	.05 (.30)	–.25 (.56)	.16 (.45)	–.02 (.29)	–.08 (.30)	.09 (.31)	–.07 (.26)
Cycle 2	–.01 (.39)	–.36 (.52)	.20 (.49)	–.09 (.40)	–.03 (.39)	.02 (.41)	.03 (.32)
Cycle 3	.05 (.48)	–.18 (.45)	.20 (.41)	.02 (.36)	–.01 (.40)	.07 (.47)	.06 (.40)

Cycle = study–test cycle; "Old" = cue was "old"; "New" = cue was "new"; Collapsed = responses for each cue were collapsed according to source.

The data of Experiment 2 confirm the explanation offered in Experiment 1. The initial study demonstrated that observers could benefit from an entirely anonymous source of recommendations even without being given performance feedback. The observed pattern suggested a low-confidence outsourcing strategy whereby they deferred to the external source's recommendation on trials of low subjective confidence (i.e., perceived guessing). This netted a benefit from the reliable source and no appreciable cost from a source that was wholly random, because low-confidence performance is near chance. The data also demonstrated that participants were similarly influenced by the reliable and unreliable sources, because the criterion shifts were similar in response to recommendations from both. Experiment 2 demonstrated that with an unreliable source that was in fact antipredictive, participants began to differentially rely on the two sources, such that they were more influenced by the reliable than by the unreliable source. Nonetheless, they failed to detect the unreliable source as antipredictive in an absolute sense. Although they lessened their reliance on it, they nonetheless treated its recommendations as generally valid.

Subjective questionnaire data The two experiments appeared to differ in the degrees to which the sources influenced responding, as indexed by criterion measures. Here we briefly consider whether these differences were also manifest in explicit awareness, as indexed by the posttest questionnaire given in both experiments. Question #4 of this instrument asked whether participants found one of the sources more reliable when they were presented simultaneously and their answers differed. For this question, the participants were to select "Source A," "Source B," or "neither source" as more accurate. Even though this question concerns primarily the double-cue condition, it should convey whether or not observers perceived an overall reliability difference between sources. In Experiment 1, 57% of respondents identified the correct source, none selected the incorrect source, and the remainder, 43%, believed the sources were similarly reliable. In Experiment 2, 73% of the respondents correctly identified one source as more reliable than the other, 3% (1 participant) incorrectly selected the unreliable/antipredictive source as more reliable, and 24% reported that neither source was more reliable. Although the tendency to correctly select the more reliable source numerically increased across experiments, this difference was not statistically significant ($p > .26$).

General discussion

The experiments reported here have identified several important patterns in the way participants use anonymous cues during recognition judgments. We consider these below.

The benefits of "conformity"

The goal of the typical memory conformity experiment is to show how people are negatively influenced by external social sources (Allan & Gabbert, 2008; Axmacher et al., 2010; Meade & Roediger, 2002; Reysen, 2005; Schneider & Watkins, 1996; Walther et al., 2002; Wright et al., 2008, 2000). Given this, the manipulations usually promote situations in which the memory performance is impaired by misinformation purposefully given by confederates. Most of these studies aim to provide data relevant to eyewitness situations, such as lineups, for example. In these cases, an incorrect memory judgment may result in catastrophic outcomes, such as false imprisonment or the inappropriate release of a dangerous criminal.

However, the focus of memory conformity experiments on such situations arguably obscures the fact that the use of external social cues in recognition should generally be viewed as a normatively beneficial strategy, and perhaps one that has been heavily evolutionarily favored. This is not to say that there are not circumstances, such as eyewitness testimony, in which errors inflated by reliance on external sources are not extremely troubling. Under such circumstances it would clearly be socially ideal if observers could entirely disregard all external sources of information and rely solely on internal memory evidence. However, these situations are not typical. Furthermore, from a functionalist

and evolutionary point of view, it remains the case that the use of external cues to bias judgments would still be beneficial even in such high-stakes situations. This becomes easier to appreciate when one considers that the memory reports of nondeceptive others clearly represent potentially useful information. As noted in the introduction, this information is not sufficient to be formally incorporated in a Bayesian fashion, because the reliability of the source is entirely unknown, rendering use of the source as a prior probability impossible.

Nonetheless, observers appear to naturally default to external sources when they experience high subjective uncertainty based on their internal evidence, a strategy we call *low-confidence outsourcing*. In short, they "conform" when they have no internal basis not to. Since it would be extremely rare to encounter an external source with no memory skill whatsoever, and since there is no reason to suspect that an anonymous source (particularly in the present paradigm) would be systematically worse than chance, this strategy arguably represents the best long-term strategy in this situation. Virtually all encountered sources will be well above chance, and hence using their reports during periods of high subjective uncertainty should almost always improve net outcomes, even though the approach is not ideal in a Bayesian sense because the problem is insufficiently specified.

Critically, we are not suggesting that observers do not take source reliability into account when it is made explicitly available, nor that they cannot learn a source's long-term reliability with prior experience and environmental feedback. Instead, the present data demonstrate that this valuable information (source reliability) is not necessary for achieving a benefit.

The limitations of establishing source reliability

Although the data suggest that observers can use the cues provided by unknown sources to their benefit (even without feedback or explicitly provided information about the sources' reliabilities), they also demonstrate severe limitations in the capacity to establish the relative reliability of external sources differing considerably in their utility (Exp. 1) or to detect a particular source as deceptive or systematically in error (Exp. 2) under these conditions. Experiment 1 demonstrated that a difference of 75% versus 50% reliability across sources appeared to largely go unnoticed by observers. It was not until the 75% reliability of the reliable source was mirrored by a 25% reliability from the unreliable source in Experiment 2 that observers clearly differentially relied upon the two. Nonetheless, they still continued to treat the unreliable source as though it provided generally valid responses (i.e., its performance was above chance). Thus, even when capable of making a relative distinction between the sources in Experiment 2, observers were still clearly unable to absolutely identify the unreliable source as in fact deceptive. This might strike some as quite remarkable; however, there are multiple factors that presumably contribute to this inability.

First, observers are not given explicit feedback regarding trial outcomes, and hence there is no registrable punishment for inappropriately using the unreliable source. Presumably such feedback-based learning is precisely how one typically learns to distrust the recommendations of others with a track record of unreliability. It remains, however, an open question for further research whether feedback, even under the conditions of Experiment 2, would lead observers to appropriately take a contrarian stance in light of an antipredictive source or whether, instead, they would simply completely discount that source of information. The latter approach would of course be nonoptimal, in the sense that an antipredictive source can provide just as much information as a predictive source. We suspect, however, that such a stance is not the natural outcome of feedback learning, and that instead behavior patterns would more closely resemble a probability-matching process (Herrnstein, 1961), whereby the differential tendency to rely on the two sources tracks their relative probabilities of reinforcement.

Second, as noted and illustrated in Experiment 1, if the observer uses either agreement or disagreement with the external agent as a proxy for its reliability (as is arguably necessary in the absence of feedback), he or she faces a considerable cognitive hurdle because, for any local or limited run of trials, it will often be the case that a source performing at chance will nonetheless frequently agree with the observer's correct, confidently held conclusions. In the absence of feedback, and given the capacity constraints of working memory in combination with random clustering, it is easy to understand how the relative utilities of the two sources could remain obscure. It is also important to note that there was little a priori reason for observers to suspect that a source might be systematically incorrect or deceptive because they were led to believe that they were viewing the responses of other students who completed the test in isolation.

The social nature of conformity effects

Finally, in contrast to typical memory conformity experiments in which confederates are used to create realistic and sometimes interactive social situations, here we show robust effects of cues in the complete absence of social interaction. The study of cue influences during memory without social interaction is actually rare (but see Betz, Skowronski, & Ostrom, 1996; Meade & Roediger, 2002). Additionally, even when social interaction is absent during

final recognition testing, social influence is often introduced earlier in the paradigm. For example, in Meade and Roediger's Experiment 4, observers viewed scenes and then either took turns recalling the scenes with a misleading confederate or took turns recalling the scenes and reading the misleading responses from fictional participants. After these initial recall experiences, observers performed a final recall and source recognition test in isolation. Thus, cues from confederates or fictional participants were given during the initial recall phase with the intention of creating false memories, while no cues were provided during the later source recognition test. This experiment revealed that the virtual and the actual confederates were equally influential on the final recall test, even though the virtual confederate was not equally influential in the source recognition test. Here, in contrast, we show that external cues directly preceding each memory probe in a standard recognition test robustly influence participants' responses, despite the absence of any actual social interaction. Thus, while social interaction may play a considerable role in conformity of judgments, it is clearly not a prerequisite for demonstrating so-called *memory conformity effects* when external cues precede the memory probes in a standard recognition memory paradigm.

Therefore, we suggest that memory conformity effects in large part reflect a beneficial decision strategy whereby participants incorporate external environmental cues about an item's memory status into their own judgments when internal evidence leads to high subjective uncertainty. We have no doubt that social factors could amplify this tendency, but the present data suggest that social interaction per se is not the fundamental aspect of conformity patterns for recognition memory judgments. From this perspective, the power of Asch's (1955) experiments and of the follow-up research in this area (e.g., Bahrami, Olsen, Latham, Roepstorff, Rees, and Frith 2010; Bond & Smith, 1996) is not a demonstration that observers bend to external cues when rendering judgments, but a demonstration that they do so even when their internal evidence is presumably unambiguous. In contrast, judgments of recognition often do not rely on unambiguous evidence, and given this, reliance on external cues is demonstrably generally beneficial.

However, the present explicit-mnemonic-cueing paradigm may be also useful for more socially informed questions. For example, only slight modification would be necessary to present the cues as originating from particular social groups, and doing so would enable one to examine whether cue influence was or was not sensitive to such factors. Alternatively, in light of the eyewitness testimony literature, one could examine whether individuals can in fact resist being influenced by external cues based on various factors. For example, if participants were appropriately informed of the cue reliabilities in Experiment 1 and were provided incentive not to utilize those cues in their reports, could they do so? We are currently researching these types of interesting questions.

Conclusions

The present findings demonstrate the benefits and limitations of the approach that observers take to incorporating external cues from anonymous sources into their recognition judgments. Overall, the data suggest a successful strategy whereby observers defer to external sources selectively under conditions of high subjective uncertainty. Prior research has repeatedly demonstrated that under standard explicit recognition tasks, subjective reports of guessing often correspond to performance near chance (Gardiner et al., 2002), and this finding was confirmed here for low-confidence reporting. Given this, the strategy is quite appropriate under the present conditions because it is rarely the case that one would encounter a source that reliably fell below chance performance (i.e., was deceptive), particularly if one is viewing the responses of prior fellow students completing an identical test. Despite having no explicit information about the sources' actual reliability and no performance feedback, participants in Experiment 1 nonetheless benefited from a reliable source and, as important, did not suffer at the hands of a source that was wholly random. Thus, the approach would provide a net benefit across an enormous range of encounters of this type. Nonetheless, the data also clearly demonstrate that the low-confidence outsourcing strategy also appears to operate when sources are of questionable reliability, and indeed Experiment 2 showed that this strategy continued to operate even when the unreliable source was deceptive or systematically incorrect. This likely reflects the fact that the detection of sources as deceptive or systematically in error is difficult in the absence of feedback and given the strong confirmatory biases observers tend to adopt in decision tasks. Of course, given the rarity with which others presumably try to actively deceive us about our own memories, the present low-confidence outsourcing strategy should hold generally. Thus, rather than characterizing the findings as a negative consequence of social conformity, we would suggest that they reflect an ingenious yet simple metacognitive strategy. Whether or not participants can "shut off" this decision strategy when given reasons to suspect external sources may be deceptive, or when the use of such sources is socially inappropriate, remains an interesting question for future study.

Author note This research was supported by National Institutes of Health Grant MH07398.

References

Allan, K., & Gabbert, F. (2008). I still think it was a banana: memorable "lies" and forgettable "truths. *Acta Psychologica, 127,* 299–308. doi:10.1016/j.actpsy.2007.06.001.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193,* 31–35. doi:10.1038/scientificamerican1155-31.

Axmacher, N., Gossen, A., Elger, C. E., & Fell, J. (2010). Graded effects of social conformity on recognition memory. *PLoS ONE, 5,* e9270. doi:10.1371/journal.pone.0009270.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329,* 1081–1085. doi:10.1126/science.1185718.

Baron, R. S., Vandello, J. A., & Brunsman, B. (1996). The forgotten variable in conformity research: impact of task importance on social influence. *Journal of Personality and Social Psychology, 71,* 915–927. doi:10.1037/0022-3514.71.5.915.

Betz, A. L., Skowronski, J. J., & Ostrom, T. M. (1996). Shared realities: social influence and stimulus memory. *Social Cognition, 14,* 113–140. doi:10.1521/soco.1996.14.2.113.

Bond, R., & Smith, P. B. (1996). Culture and conformity: a meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119,* 111–137. doi:10.1037/0033-2909.119.1.111.

Gabbert, F., Memon, A., & Wright, D. B. (2007). I saw it for longer than you: the relationship between perceived encoding duration and memory conformity. *Acta Psychologica, 124,* 319–331. doi:10.1016/j.actpsy.2006.03.009.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: a meta-analysis of remember, know, and guess responses. *Memory, 10,* 83–98. doi:10.1080/09658210143000281.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of memory recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1176–1195. doi:10.1037/0278-7393.30.6.1176.

Herrnstein, R. J. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4,* 267–272. doi:10.1901/jeab.1961.4-267.

Hoffman, H. G., Granhag, P. A., Kwong See, S. T., & Loftus, E. F. (2001). Social influences on reality-monitoring decisions. *Memory & Cognition, 29,* 394–404. doi:10.3758/BF03196390.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Meade, M. L., & Roediger, H. L., III. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30,* 995–1009. doi:10.3758/BF03194318.

O'Connor, A. R., Han, S., & Dobbins, I. G. (2010). The inferior parietal lobule and recognition memory: expectancy violation or successful retrieval? *Journal of Neuroscience, 30,* 2924–2934. doi:10.1523/JNEUROSCI.4225-09.2010.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General, 109,* 160–174. doi:10.1037/0096-3445.109.2.160.

Reysen, M. B. (2005). The effects of conformity on recognition judgments. *Memory, 13,* 87–94. doi:10.1080/09658210344000602.

Schneider, D. M., & Watkins, M. J. (1996). Response conformity in recognition testing. *Psychonomic Bulletin & Review, 3,* 481–485. doi:10.3758/BF03214550.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50. doi:10.1037/0096-3445.117.1.34.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104–132. doi:10.1037/0033-295X.114.1.104.

Walther, E., Bless, H., Strack, F., Rackstraw, P., Wagner, D., & Werth, L. (2002). Conformity effects in memory as a function of group size, dissenter and uncertainty. *Applied Cognitive Psychology, 16,* 793–810. doi:10.1002/acp.828.

Wright, D. B., Gabbert, F., Memon, A., & London, K. (2008). Changing the criterion for memory conformity in free recall and recognition. *Memory, 16,* 137–148. doi:10.1080/09658210701836174.

Wright, D. B., Self, G., & Justice, C. (2000). Memory conformity: exploring misinformation effects when presented by another person. *British Journal of Psychology, 91,* 189–202. doi:10.1348/000712600161781.