Psychon Bull Rev (2015) 22:509–516
DOI 10.3758/s13423-014-0703-5

BRIEF REPORT

The role of mnemonic processes in pure-target and pure-foil recognition memory

Gregory J. Koop · Amy H. Criss · Kenneth J. Malmberg

Abstract Surprisingly, response patterns in a recognition memory test are very similar regardless of whether the test list contains both targets and foils or just one class of items. To better understand these effects, we evaluate performance over the course of testing. Output interference (OI) is the decrease in performance across test trials due to an increase in noise caused by encoded test items. Critically, OI is predicted on pure lists if the mnemonic evidence for each test item is evaluated. In two experiments, participants received accurate feedback, no feedback, or random feedback that was unrelated to the response on each test trial and pure or standard test lists. When no feedback was provided, performance was nearly identical for standard and pure test lists, replicating previous findings. Only in the presence of accurate feedback were participants able to successfully adapt to pure list environments and improve their accuracy. Critically, OI was observed, demonstrating that participants continued to evaluate mnemonic evidence even in pure list conditions. We discuss the implication of these data for models of memory.

Keywords Output interference · Feedback · Test composition · Criterion shifts

Recognition memory is often examined by asking participants to study a list of words, followed by a series of test trials in which words are classified as studied targets or unstudied foils. Signal detection theory (SDT; Green & Swets, 1966) is often used to measure the accuracy and bias of these decisions. Such an analysis necessitates the inclusion of both targets and foils during the test, a practice that is common to nearly all studies in the literature. However, in some experiments, the proportion of targets and foil trials is manipulated in order to induce changes in bias, and the probability of calling any test item "studied" increases slightly as the proportion of targets increases (e.g., Criss, 2009; Healy & Kubovy, 1978; Ratcliff, Sheu, & Gronlund, 1992; Rotello, Macmillan, Hicks, & Hautus, 2006). In an extreme example, Cox and Dobbins (2011; see also Wallace, 1982) used test lists that were distractor-free or target-free (i.e., pure test lists) in one condition, whereas in the other condition, the test lists consisted of an equal number of target and foil trials (i.e., standard test lists). In light of prior findings showing shifts in bias consistent with the predictions of SDT, the startling result reported by Cox and Dobbins was that hit rates (HRs) were virtually identical for standard and pure test lists, whereas false alarm rates (FARs) were very slightly higher for pure foil test lists than for standard test lists. The near insensitivity to pure test lists suggests that participants are not adopting a flexible criterion responsive to the specific experimental paradigm. Furthermore, these data are not consistent with a fully informed likelihood model of memory where the cognitive system takes into account the expected (or actual) target and foil distributions and adjusts the decision rule accordingly, often to optimize performance (e.g., Glanzer, Hilford, & Maloney, 2009; Starns, White, & Ratcliff, 2010; Treisman & Williams, 1984).

Why might participants be almost completely insensitive to test lists consisting entirely of targets or entirely of foils? Two accounts, one described in terms of reinforcement learning history (Mickes, Hwe, Wais, & Wixted, 2011) and the other in terms of preexperimental priors (Turner, Van Zandt, & Brown, 2011), suggest that a lifetime of recognition experience creates a well-formed understanding of memory strength such that participants do not deviate from these assumed distributions unless they are provided with an indication that these

---

G. J. Koop (✉) · A. H. Criss
Department of Psychology, Syracuse University, Syracuse, NY 13244, USA
e-mail: gregory.koop@emu.edu

K. J. Malmberg
University of South Florida, Tampa, FL, USA

distributions no longer apply. In short, these models accommodate complete insensitivity to both pure foil and pure target lists in the absence of external feedback. However, it is unclear whether this prediction is accurate, because there appears to be a slight increase in false alarms for pure foil test lists. The key challenge is to explain why the shift in bias occurs only for pure foil lists and not for pure target lists. Cox and Dobbins (2011) developed the *motivated-to-recognize* hypothesis, proposing that participants expect to remember some proportion of items and are biased to respond in a way that meets this expectation. In a pure target test, this expectation is met, and participants continue to give "studied" responses at the typical rate, and without any indication of poor performance, they do not update this response pattern. In a pure foil test, however, participants are unable to meet their expected allowance of "studied" responses and, therefore, adopt a slightly more liberal criterion, which presumably occurs during the course of recognition testing.

Because the pure list data are so compelling and inconsistent with a flexible response criterion—a fundamental assumption of most memory models—it is critical to pursue replication and a better understanding of performance in pure test lists. We leverage output interference (OI), the finding that accuracy declines over testing, to better understand pure test lists. OI in recognition is commonly observed in the form of a steep decrease in HRs and a more variable pattern of FARs, but always such that overall accuracy decreases as test trial increases (Annis, Malmberg, Criss, & Shiffrin, 2013; Criss, Malmberg, & Shiffrin, 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson 1975). OI has been modeled as the result of encoding during test (Criss et al., 2011). Specifically, OI is well characterized by a model with the following properties. During the course of testing, on a trial where an item is judged to have been studied, the best matching episodic trace is updated with additional information. When an item is judged to be new during the memory test, a new episodic trace is stored. These two factors, strengthening existing traces and storing new traces, both cause a decrease in HR for subsequently tested target items. The two factors have opposing effects on the FAR, with the former decreasing the FAR via differentiation and the latter increasing the FAR (see Criss & Koop, in press, for a review). Thus, the end result is often no change in the FAR across test trials, but the exact pattern of FARs depends on the details of performance (for full details, see Criss et al., 2011). Note that the model operates in the same way for pure and standard test lists (e.g., the model only takes into account the response on a given test trial to modulate storage during test) and OI should, therefore, be present for both types of lists. In summary, we broadly characterize OI as a decrease in HR accompanied by a roughly flat FAR. The critical prediction is that if individuals rely on mnemonic evidence at test, OI should be present even in pure lists. So long as the test item is compared with the contents of episodic memory and the contents of episodic memory include information about prior test items, OI should be observed.

Experiments 1 and 2

We compared pure tests (composed of all targets or all foils) with standard tests (composed of an equal number of targets and foils), and we systematically crossed this manipulation with different types of feedback (none, accurate, and random) between participants to measure adjustments to the decision criterion and changes in accuracy across test trials. Participants were not explicitly informed about either of these manipulations. Accurate feedback was originally included in OI experiments to address the criticism that OI is the result of decreasing motivation as the test proceeds. Despite demonstrating clear patterns of OI and release from OI regardless of the presence or absence of feedback (Criss et al., 2011; Malmberg et al., 2012), we included both accurate and random feedback conditions in these experiments to further disentangle the effects of accurate feedback from the motivational effects of any feedback whatsoever.

Recent models have proposed that feedback plays a critical role in performance by allowing participants to create representations of the information distributions on which the decisions are made (Turner et al., 2011). However, within the recognition memory literature, the effects of accurate feedback are equivocal. Many studies have shown that accurate feedback has little, if any, impact on recognition accuracy (Criss et al., 2011; Han & Dobbins, 2008, Experiment 1; Kantner & Lindsay, 2010; Malmberg & Annis, 2012). However, feedback does, at least in some cases, change response bias (Kantner & Lindsay, 2010; Starns, White, & Ratcliff, 2010). Additionally, in test scenarios with multiple criteria (e.g., confidence judgments), corrective feedback may affect the spread of criteria. In one such case, relative to no-feedback conditions, the most conservative (highest confidence) criteria became more conservative, whereas the most liberal (lowest confidence) criteria became more liberal (Mickes et al., 2011). Although feedback has been studied in many standard testing scenarios, the effect of feedback on pure test lists remains an open question. Experiments 1 and 2 are typical single-item recognition tests with a study list followed by a test list and instructions to judge whether the item is studied or new as accurately as possible. The only difference between the experiments was the manipulation of normative word frequency (WF) in Experiment 2.

Method

Participants

Two hundred ninety-four undergraduates participated in Experiment 1, and 269 participated in Experiment 2. All participants were drawn from the Syracuse University research participation pool and received credit in partial fulfillment of course requirements.

Stimulus materials

The word pool for Experiment 1 consisted of 800 high normative frequency (HF) words between 4 and 11 letters in length ranging between 9 and 13 log frequency ($M = 10.46$) in the Hyperspace Analog to Language corpus (HAL; Balota et al., 2007). Experiment 2 also included 800 low normative frequency (LF) words between 4 and 11 letters in length ranging between 3.5 and 6 log frequency ($M = 5.22$) in the HAL corpus. For each experiment, a subset of the respective word pool was randomly assigned to each condition for each participant.

Design and procedure

Experiment 1 was a 3 (feedback: none, accurate, or random) × 3 (test composition: pure target, pure foils, or standard half-target/half-foil) between-participants design where each participant performed a single study–test cycle. Participants were not informed about test composition or feedback manipulations. Experiment 2 included WF as a factor, with equal numbers of HF and LF words in every condition. Each study list consisted of 120 words presented for 1 s, with a 1-s blank screen separating trials. A 45-s running arithmetic task separated study and test, with additional time for self-paced reading of instructions.

Each test list consisted of 120 self-paced trials. Participants were informed that this was a memory test and they should judge whether or not they saw each item in the study list. After indicating a “studied” or “unstudied” response, participants received feedback for 100 ms, followed by a 100-ms blank screen. In the feedback-present conditions, participants saw the word “wrong” or “correct” in the middle of the screen. In the accurate condition, this feedback was truthful, whereas in the random condition, the feedback was randomly selected with equal probability. In the no-feedback condition, a blank screen was displayed for the duration of the feedback window. In the standard condition, half of the test items were targets and half were foils, whereas in the pure condition, all of the test items were either targets or foils.

Results and discussion

Because the pure list condition was manipulated entirely between participants, half of the pure list participants did not have a FAR (the pure targets condition), whereas the other half did not have a HR (the pure foils condition). Thus, on all figures, the pure list conditions represent the performance of two different groups of participants, whereas plots of the HR and FAR for standard list conditions reflect a single group of participants. Statistical significance was assessed against a threshold of $p = .05$, with Greenhouse–Geisser adjusted degrees of freedom where necessary. When evaluating OI, we report trend analyses to better characterize the form of the differences between test blocks. Linear trends are the form typically observed for OI and will be the focus of our analyses.

Output interference and test composition in the absence of feedback

We first examined the conditions in Experiments 1 and 2 (pure and standard lists with no feedback) that most closely replicated the original design of Cox and Dobbins (2011). Note that the only critical difference is that the participants in our experiments were given typical instructions to endorse items they remembered from the study list and reject others, whereas Cox and Dobbins emphasized subjective feelings of familiarity, with no emphasis on accuracy. As is readily apparent in the first column of Table 1, neither the HR, $t(76) = 0.19$, $p = .847$, nor the FAR, $t(70) = 0.26$, $p = .798$, differed between

Table 1 Proportion *old* for each condition in Experiments 1 and 2

	Feedback		
Test CONDITION	None *M* (*SE*)	Accurate *M* (*SE*)	Random *M* (*SE*)
Experiment 1			
Pure targets	.55 (.02)	.88 (.03)	.56 (.02)
Standard targets	.55 (.03)	.64 (.03)	.62 (.03)
Pure foils	.31 (.02)	.05 (.02)	.49 (.02)
Standard foils	.30 (.02)	.37 (.03)	.40 (.03)
Experiment 2: High-Frequency Words			
Pure targets	.52 (.04)	.92 (.01)	.53 (.02)
Standard targets	.55 (.03)	.64 (.02)	.61 (.03)
Pure foils	.24 (.03)	.03 (.01)	.51 (.02)
Standard foils	.31 (.03)	.32 (.03)	.37 (.03)
Experiment 2: Low-Frequency Words			
Pure targets	.62 (.03)	.90 (.02)	.66 (.02)
Standard targets	.66 (.03)	.72 (.03)	.69 (.03)
Pure foils	.20 (.03)	.05 (.01)	.41 (.02)
Standard foils	.25 (.02)	.26 (.03)	.32 (.03)

pure and standard test lists for Experiment 1, indicating insensitivity to test composition without feedback. In Experiment 2, 2 (WF) × 2 (test composition) mixed ANOVAs similarly failed to show a difference between test compositions [HR, $F(1, 56) = 0.69$, $p = .409$; FAR, $F(1, 62) = 2.97$, $p = .09$] but did show the typical finding that performance was better for LF words than for HF words [HR, $F(1, 56) = 20.36$, $p < .001$; FAR, $F(1, 62) = 9.99$, $p = .002$].

To evaluate OI, we divided the test lists into six blocks of 20 trials. All effects reported in the previous paragraph remain unchanged by the inclusion of test block as a factor; therefore, we report only the main effects and interactions associated with test block. OI was observed in both experiments. As is typically the case, HRs decreased, whereas FARs remained flat or only slightly changed. The no-feedback conditions of Experiment 1 showed a typical decrease in HR across blocks, in the form of a linear trend, $F(1, 76) = 39.22$, $p < .001$, and no linear trend for the FAR, $F(1, 70) = .002$, $p = .966$ (shown in Fig. 1). Similarly, the no-feedback conditions in Experiment 2 demonstrated a decrease in HR across test block in the form of a linear trend, $F(1,5\ 6) = 20.14$, $p < .001$ (see Figs. 2 and 3). Unlike in Experiment 1, there was a block × test composition interaction, $F(4.20, 260.55) = 3.16$, $p = .013$, because the pattern of FARs slightly differed for the pure and standard test lists.

We replicated one of the two main findings from Cox and Dobbins (2011). Namely, HRs were virtually identical between standard lists and pure target lists, and this pattern persists in the presence of OI. However, we failed to fully support the *motivated-to-recognize* account because participants did not become more liberal in pure foils lists, relative to standard lists. Observation of a typical pattern of OI in the pure test lists indicates that participants continue to evaluate the quality of mnemonic evidence, which decreases as information is added to memory during testing.

Output interference and accurate feedback during pure test lists

Having shown virtually identical performance for pure and standard test lists in the absence of feedback, we now turn to an evaluation of pure test lists when feedback was provided. From the figures, it is clear that accurate feedback improves performance in pure lists. Although this finding would not be expected on the basis of prior literature examining pure lists or feedback in isolation, many readers will find this result somewhat intuitive. However, the exact nature of this improvement is still of interest, especially regarding the claim that participants continue to evaluate evidence over the course of the test. If so, HR should decrease as a result of OI, perhaps after an initial increase due to accurate feedback. In other words, HRs should show a quadratic trend. In fact, the data show just such a pattern.

In Experiment 1, accurate feedback produced a quadratic trend for HR, $F(1, 24) = 15.42$, $p = .002$, due to a substantial increase in performance during the first two test blocks, followed by a decline in HR across remaining test blocks (see Fig. 1). As further evidence, we conducted a simple comparison of block 2 with block 6, which demonstrated that performance decreased in accordance with OI, $t(24) = 2.28$, $p = .032$. FARs decreased over the course of testing, evidenced by a linear contrast, $F(1, 39) = 19.61$, $p < .001$, with a precipitous early decrease before leveling off for the later test blocks. Likewise, in Experiment 2 (see Figs. 2 and 3), accurate feedback produced a sharp increase in HR, which was

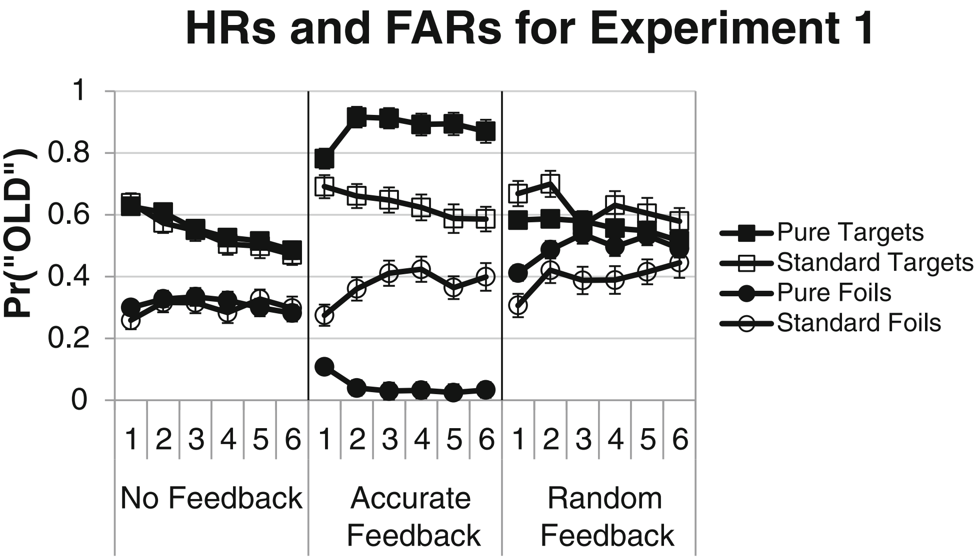

Fig. 1 Proportion “OLD” across test blocks (1–6) for each feedback condition in Experiment 1. Filled points indicate pure tests, and empty points represent standard tests. Bars = ±1 *SE*

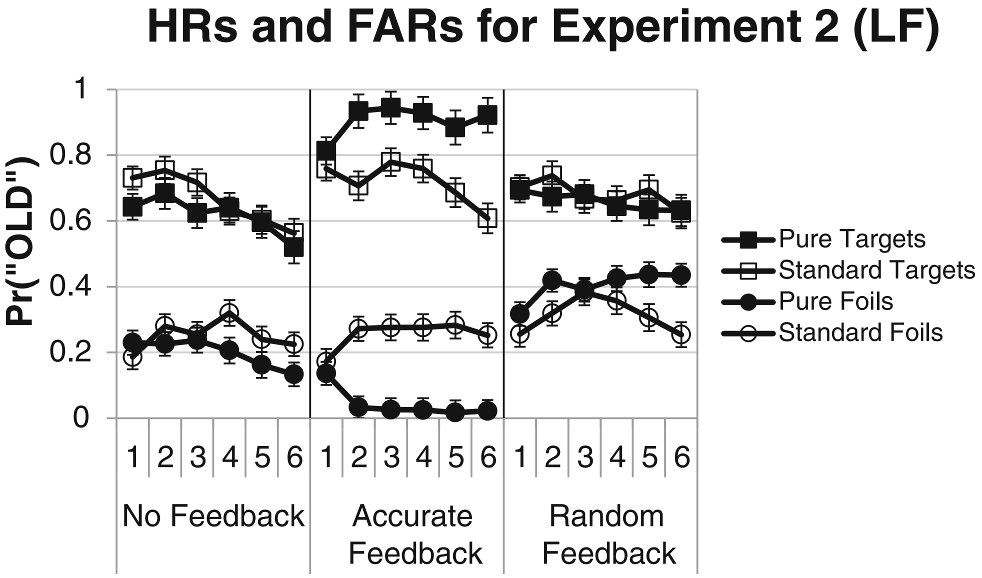

Fig. 2 Proportion "OLD" for low-frequency words across test blocks (1–6) for each feedback condition in Experiment 2. Filled points indicate pure tests, and empty points represent standard tests. Bars = ±1 *SE*

followed by a very slight decrease over remaining blocks, resulting in a quadratic contrast for HRs, $F(1, 21) = 50.93$, $p < .001$, whereas FARs decreased over the course of testing, as confirmed by a linear contrast, $F(1, 36) = 18.58$, $p < .001$. The pattern of OI was consistent for LF and HF words in Experiment 2 [i.e., no interaction between test block and WF for HRs, $F(5, 130) = 1.30$, $p = .269$, or FARs, $F(1.94, 69.70) = 2.55$, $p = .087$].

The data suggest that participants do not simply respond "yes" on every trial of a pure target test list or "no" on every trial of a pure foil test list, despite being given feedback on each trial indicating the correct response. This is perhaps just as surprising as the fact that participants in pure lists without feedback provide responses that are nearly identical to those given in a standard test list. The data from pure test lists with accurate feedback are consistent with the idea that participants continue to evaluate the evidence provided by the test cue. Whereas the FAR is not diagnostic, the decrease in HR strongly suggests that test items continue to be compared with the contents of episodic memory and that the evidence decreases as information is added to memory over the course of testing.

One simple explanation of these data is that one subgroup of participants ignored the feedback and continued to treat the task as a typical memory test, whereas a second (possibly larger) group realized that they were in a pure list and subsequently responded accurately on every trial without evaluating mnemonic evidence. To briefly address this concern, we looked at individual-level data. Overall, individuals showed a pattern similar to the aggregate data; for example, only 36 % of participants in Experiment 1 and 59 % of participants in Experiment 2 are 100 % accurate during the final test block, suggesting that there are not subgroups whose combined data appear to be OI in pure lists.[1]

Output interference and random feedback during pure test lists

For pure lists, the effects of random feedback clearly differed from those of accurate feedback, which indicates that accurate information, not simply the presence of feedback, is necessary for individuals to capitalize on extreme test list distributions. In response to random feedback, HRs decreased and FARs increased such that they approached 0.50. Linear trends for the HRs were observed in Experiment 1, $F(1, 37) = 4.80$, $p = .035$, but the same numeric pattern was not significant in Experiment 2, $F(1, 26) = 1.23$, $p = .278$. FARs showed an increasing linear trend in both Experiment 1, $F(1, 30) = 7.40$, $p = .011$, and Experiment 2, $F(1, 33) = 6.17$, $p = .018$. One possible explanation for this effect is the *motivated-to-recognize* account. It is possible, for example, that participants' general inclination to identify some subset of items as "old" is occasionally reinforced (incorrectly) by feedback, which therefore supports the adoption of an increasingly liberal response strategy.[2] Another possibility is that the increase in FAR and decrease in HR reflect OI.

---

[1] Details including plots of individual-participant analyses are available by request.

[2] We thank Ian Dobbins for this suggestion.

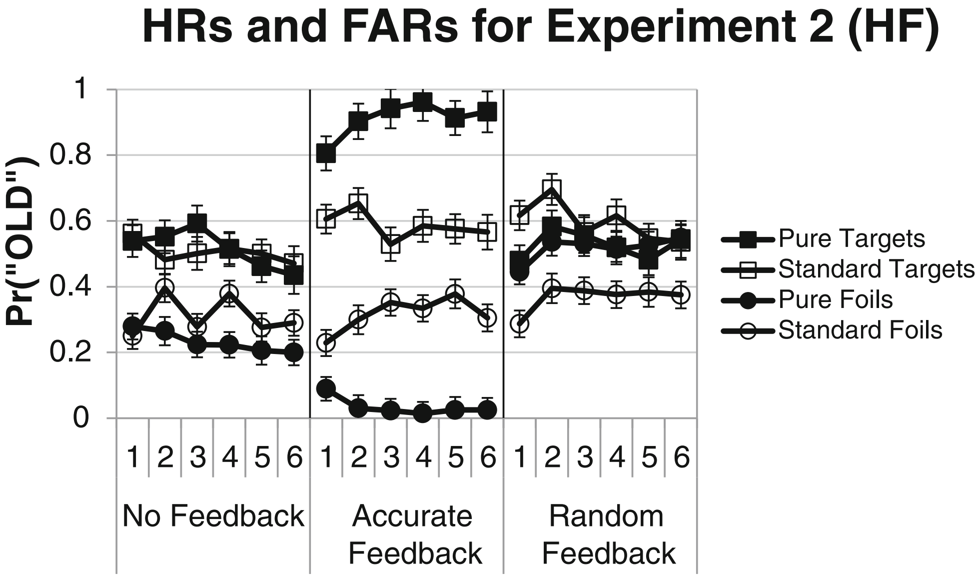

Fig. 3 Proportion "OLD" for high-frequency words across test blocks (1–6) for each feedback condition in Experiment 2. Filled points indicate pure tests, and empty points represent standard tests. Bars = ±1 *SE*

Output interference and feedback during standard test lists

In our final set of analyses, we examine the effects of OI and feedback in the standard test list conditions.[3] Table 2 provides bias and discriminability measures from SDT. In Experiment 1, feedback had an effect on participants' criteria, *F*(2, 81) = 5.17, *p* = .008. Bonferroni-corrected post hoc comparisons showed that participants were more conservative in the no-feedback condition, relative to the accurate, *t*(59) = 2.64, *p* = .030, and random , *t*(55) = 2.72, *p* = .024, feedback conditions. However, the *content* of feedback did not impact *C*, since accurate and random feedback conditions did not differ, *t*(44) = 0.20, *p* > .05. Although there was not an overall effect of feedback, *F*(1, 88) = 2.02, *p* = .139, in Experiment 2, the same descriptive pattern of data emerged. Participants tended toward less conservative responding—that is, saying "studied" more often—when provided with feedback, regardless of whether that feedback was accurate or random. Discriminability was not affected by feedback in Experiment 1, *F*(2, 79) = 0.53, *p* = .532. Experiment 2 showed an effect of feedback, *F*(2, 88) = 3.18, *p* = .047, which did not interact with WF, *F*(2, 88) = 0.18, *p* = .834. However, no post hoc comparisons between feedback conditions survived Bonferroni correction. Taken together, the results from both experiments suggest that if feedback affects discriminability in standard lists, its impact is modest.

OI was observed for all standard test lists in Experiments 1 and 2 in the form of decreasing HRs and slightly increasing or flat FARs and the pattern of OI did not interact with the feedback condition [Experiment 1: linear trend for HRs, *F*(1, 79) = 21.00, *p* < .001, and FAR, *F*(1, 79) = 9.49, *p* = .003; Experiment 2: linear trend for HR, *F*(1, 88) = 19.21, *p* < .001, but a nonsignificant linear trend for FAR, *F*(1, 88) = 2.27, *p* =.136]. The no-feedback and accurate feedback conditions replicate prior work (e.g., Criss et al., 2011), and the random feedback condition suggests that the content of the feedback does not affect the extent to which people evaluate the contents of episodic memory and encode information into episodic memory during standard recognition tests.

General discussion

Collectively, these analyses replicate and extend the finding that participants respond in largely the same way regardless of whether test lists consist of all targets, all foils, or a standard mix of both targets and foils. Feedback, whether accurate or random, results in a more liberal response bias for standard test lists than does no feedback. In pure lists, providing accurate feedback increases HR and decreases FAR, as intuition would suggest, while random feedback drives performance toward responding "old" at rate of .50 for both pure target and pure foil test lists. Critically, OI is present in all conditions in the form of a decrease in HR and a variable pattern of FAR across test block.

The similarity in performance between pure and standard test lists is quite surprising because it contrasts with the long-known finding of a systematic change in response bias with changes in target base rate. However, a more careful review of

---

[3] Note that the analysis of SDT parameters cannot be conducted for the pure list conditions, because an individual provides either an HR or an FAR but not both.

Table 2 Discriminability (*d'*) and response bias (*C*) for standard lists in Experiments 1 and 2

Feedback	None		Accurate		Random	
	d' (*SE*)	*C* (*SE*)	*d'* (*SE*)	*C* (*SE*)	*d'* (*SE*)	*C* (*SE*)
Experiment 1	0.68 (.06)	0.22 (.06)	0.73 (.08)	−0.01 (.06)	0.61 (.08)	−0.03 (.08)
Experiment 2						
LWF	1.19 (.10)	0.15 (.07)	1.32 (.10)	0.04 (.07)	1.08 (.10)	−0.02 (.07)
HWF	0.70 (.06)	0.21 (.08)	0.89 (.06)	0.07 (.08)	0.69 (.07)	0.03 (.08)

Note. LWF, low word frequency; HWF, high word frequency

the literature shows that the findings for target base rate manipulations are smaller or absent when participants are not informed of the manipulation (e.g., Koop & Criss, 2013; Rhodes & Jacoby, 2007). In this context, the pure test list data are less surprising. The present data show that without feedback, participants do not adapt to a foil-free or target-free environment. However, when provided with feedback during test, participants adapt, not to the test composition per se, but to the composition suggested by feedback. In the random feedback condition, participants move toward responding "old" 50 % of the time in both pure lists and standard lists. In the accurate feedback condition, participants move toward responding "old" all of the time in the pure target condition and "new" all of the time in the pure foil condition. The use of the words "move toward" is intentional because, although responding does change, participants still evaluate mnemonic evidence.

One simple explanation of the finding that feedback enables participants to capitalize on pure test lists is that feedback simply made the underlying distribution explicit and participants mindlessly selected the appropriate response for the remainder of the task. However, the presence of OI contradicts this account. OI is the result of information being encoded into memory during test in the form of either storing new traces for unrecognized test items or updating an episodic trace that is a strong match for a recognized item. This newly encoded information adds noise to the mnemonic evidence for subsequently tested items, reducing performance (Criss et al., 2011). Thus, from a memory-based framework, OI should be observed regardless of decisional strategies employed by participants and persist even across pure lists, which is consistent with our data. The presence of OI in the data demonstrated that participants continued to evaluate the evidence provided by the stimulus in pure lists when feedback was withheld. When accurate feedback was provided, participants continued to show OI even after adapting to the test environment suggested by accurate feedback.

Although precise predictions about the magnitude of OI in different conditions are possible, the role of feedback has not yet been incorporated into the model. For example, Criss, Kiliç, Malmberg, and Fontaine (2014) demonstrated how the relative contributions of trace updating and the addition of new traces affect the exact pattern of OI. Kiliç (2013) modeled the relationship between encoding strength and OI. In addition, she analyzed response time distributions with the drift diffusion model (DDM). In short, the drift rate parameter of the DDM decreased in magnitude as a function of test position, further supporting the claim that interference increases over the course of testing. A promising avenue for future research is to model the role of feedback and how it relates to OI.

The data presented here indicate that participants still make recognition decisions based on mnemonic evidence when the list consists entirely of targets or entirely of foils and even when feedback is provided that reveals this pure list structure. One common assumption in the literature is that the criterion reflects the difficulty of the test situation (e.g., Hicks & Starns, 2014; Starns, White, & Ratcliff, 2012). The pure test list situation is an ideal environment in which to evaluate this claim, and the data suggest that participants do not set the criterion in response to the ease of test items. These and related assumptions should be reevaluated in light of the findings that pure and standard test lists result in similar response patterns.

Acknowledgments This material is partially based upon work supported by the National Science Foundation under grant number 0951612.

References

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1365–1376.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940.

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297–319.

Criss, A. H., Kiliç, A., Malmberg, K. J., & Fontaine, J. (2014). *Learning during study and test: A joint evaluation of list length effects and output interference*. Manuscript in preparation.

Criss, A. H., & Koop, G. J. (in press). Differentiation in episodic memory. In J. Raaijmakers, A. H. Criss, R. Goldstone, R. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A Festschrift for Richard M. Shiffrin.* Psychology Press.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin and Review, 16*(3), 431–455. doi:10.3758/PBR.16.3.431

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Krieger.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory and Cognition, 36,* 703–715.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6,* 544–553.

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*. doi:10.3758/s13421-014-0397-y

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406.

Kiliç, A. (2013). *Output interference and strength based mirror effect in recognition memory* (Doctoral dissertation). Retrieved from SURFACE Psychology-Dissertations.

Koop, G. J., & Criss, A. H. (2013). *The response dynamics of recognition memory: Sensitivity and bias*. Toronto, ONT: Poster session presented at the meeting of the Psychonomic Society.

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141,* 233–259.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23,* 115–119.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140,* 239–257.

Murdock, B. B., Jr., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In *Information processing and cognition: The Loyola symposium* (pp. 145–194). Hillsdale, NJ: Erlbaum.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition, 34,* 1598–1614.

Starns, J. J., White, C. N., & Ratcliff, R. R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63,* 18–34.

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40,* 1189–1199.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118,* 583–613.

Wallace, W. P. (1982). Distractor-free recognition tests of memory. *The American Journal of Psychology, 95,* 421–440.