A Search for Influences of Feedback on Recognition of Music, Poetry, and Art

D. Stephen Lindsay and Justin Kantner

University of Victoria

Draft of a chapter to appear in P. Higham and J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea*. Houndmills, UK: Palgrave Macmillan.

Please address correspondence to Steve Lindsay, Psychology, UVic, P.O. Box 3050 STN CSC, Victoria, BC, V8T 2M5, Canada. E-mail slindsay@uvic.ca. Phone 1-250-721-8593. Fax 1-250-721-8929.

The stream of consciousness has many tributaries. At any given moment some are dry beds, others gently burbling brooks, yet others gushing torrents. The main current is not always mindful of the sources of its flow, such that, for example, the waters of the wellspring of memory may mingle with the freshet of insight. But, to push the metaphor past the breaking point, we'd be at sea if we were completely unable to distinguish observation from expectation, reality from wish or fear, inherent ease from familiarity, etc. Thus the mind/brain needs mechanisms for monitoring (albeit imperfectly and at varying levels of specificity) the sources of influence on its own productions. Bruce Whittlesea, whose work this volume honours, developed a theoretical perspective called Selective Construction and Preservation of Experience (SCAPE). SCAPE includes several theoretical constructs and can be applied to a number of domains (e.g., aesthetic judgments), but we focus on Whittlesea's "fluency discrepancy hypothesis" in the context of recognition memory. This hypothesis holds that mental productions (thoughts, images, etc.) in response to recognition probes in particular contexts are generated in accord with the transfer-appropriate processing principle (Morris, Bransford, & Franks, 1977) and that those products are monitored and evaluated on the fly. If the fluency (i.e., ease and speed) with which a test probe is processed differs from expectations for that item in that context, the monitoring process attributes that discrepancy to some source. In Whittlesea's (2005) words, "people attribute their self-evaluation (of the fluency of their mental productions) to some source that makes sense, given those aspects of the stimuli that are salient to them given the task and context and their intuitive causal theories." Thus, for example, subjects are prone to false-alarm to new words on a memory test if those words are processed unexpectedly easily (e.g., as in the case of orthographically regular pseudowords such as "hension;" Whittlesea & Williams, 1998, 2000).

SCAPE can be described as a major elaboration of Jacoby's attribution-making approach to memory (Jacoby & Dallas, 1981; Jacoby, Kelley, & Woloshyn, 1989; Kelley & Rhodes, 2002). In a representative experiment, Jacoby and Whitehouse (1989) found that priming recognition test words created a bias to judge those words as old if the primes were presented so briefly that subjects were unaware of them. Primes presented for a longer duration had the opposite effect, creating a bias to reject the primed test word (presumably

because their fluency was over-attributed to the prime rather than to oldness).

The source monitoring (SM) framework is yet another approach to the general issue of how the mind/brain identifies the sources of its own productions (Johnson, Hashtroudi, & Lindsay, 1993; Lindsay, 2008). Most work inspired by the SM framework has addressed the issue of how people differentiate between memories from different sources (such as remembering when and where a past experience occurred, its medium of presentation and modality of perception, the actors involved in it, etc.). The core idea of the SMF is that aspects of the sources of thoughts and images are inferred from the perceptual, semantic, and affective content of those thoughts and images. For example, memory information that comes to mind from a past conversation with your friend Lee might include information about the meaning of Lee's utterance, the sound of his voice, his appearance and that of the surrounding context, your thoughts and feelings at the time, etc., all of which provide clues to various dimensions of the source of this recollection.

Jacoby's attribution-making approach, Whittlesea's SCAPE, and the SM framework all hold that source inferences are usually made quickly and without conscious deliberation via heuristic processes. People occasionally experience uncertainty as to the provenance of a mental event ("Did I lock the door when I left or did I only think about locking the door?") and they may then use conscious strategies in an effort to identify source. Such deliberative processes may or may not yield a feeling of having successfully identified aspects of source, and if they do it may or may not be justified (i.e., we can identify an aspect of source reflectively and nonetheless be wrong). But most of the time we are no more aware of these monitoring processes than we are of the inferential processes involved in, say, seeing a three-dimensional world around us. The inferences generally unfold quickly and automatically in the flow of ongoing mental processing.

Inferential attribution-making processes have been the focus of most of the research inspired by the attribution-making, SCAPE, and SM approaches. But the processes that lead mental contents to come to mind in response to cues in contexts are also crucial to memory performance and experience. All three of the approaches emphasized here assume that transfer-appropriate processing (Morris et al., 1977) drives the generation or production of mental events. Thus the way subjects think about recognition probes at test influences the mental events that are generated in response to those probes, which in turn affects inferences about their sources.

The attribution-making approach, SCAPE, and the SM framework share the key notion that the mind/brain evaluates its own products and makes inferences about their sources. A casual reading might create the impression that these theoretical perspectives posit two discrete (non-overlapping) stages: First a mental product is generated and then later it is evaluated. But production and evaluation are always going on concurrently, and the outcome of monitoring processes affects generation processes, just as the characteristics of mental productions affect the outcome of monitoring processes.

In the interest of theory development, research inspired by these ideas has typically focused on errors in the attribution process; when a variable is shown to modulate the likelihood of false attributions that finding supports the general claim that there are attribution-making processes and sheds light on the specifics of their operation. But most attributions in everyday life are accurate. Mental products that have the properties of memories usually are memories and usually are experienced as memories, those with the characteristics of perception usually are based (primarily) on perception, etc.

The attribution-making, SCAPE, and SM approaches hold that people use various heuristics to attribute mental events to source. For example, Jacoby and Whittlesea both hold that in the context of a recognition memory test people tend to be biased to attribute unexpectedly fluent processing to the use of memory and hence tend to judge such items as old, whereas in another context they might attribute fluency to well-formedness. In research on another sort of source-attribution bias, Johnson, Raye, Foley, and Foley (1981) had subjects read some words aloud and listen to another person say other words aloud; later, if subjects false alarmed to non-studied test items they were biased to attribute that item to the other person rather than themselves (the "It had to be you" effect). Presumably this reflects a (perhaps unconscious) grasp of the fact that memory tends to be better for self-generated and enacted events than for passively perceived events (Cohen, 1989; Slamecka & Graf, 1978).

How do people come by such source-attribution heuristics and biases? Are these built-in tendencies, fully available from an early age, or do they develop gradually during childhood? We do not know of any studies of children's memory informed by Jacoby's attributional approach to memory, nor by Whittlesea's SCAPE model, and these seem to us to be rich fields for future inquiry. The SM framework has inspired a fair number of child development studies (see, e.g., an edited volume by Roberts & Blade, 2000). Simplifying greatly, this research indicates that young preschoolers can do as well as adults at discriminating the sources of memories under some conditions (including, interestingly, conditions that lead to performance that is below ceiling in all age groups), but that they tend to do much more poorly than older children or adults when discriminations are particularly difficult (e.g., when candidate sources share many perceptual and/or conceptual characteristics or when delays are long).

The existence of developmental changes in SM performance suggests that experience may play a role in shaping the biases and heuristics that guide the attribution of mental events to particular sources. Indeed, such a claim is more or less explicit in Jacoby and co-authors' exposition of the attribution-making approach: It is because of the fact that use of memory facilitates processing that the system comes to be biased to interpret unexpected fluency as diagnostic of oldness. As direct support of this idea, Unkelbach (2006) demonstrated that a training experience in which experimentally manipulated high fluency was always linked to new items and low fluency to old items led to a reversal of the usual fluency effect; that is, subjects developed a bias to respond "New" to highly fluent items and "Old" to non-fluent items (see Unkelbach, 2007, for analogous findings with truth judgments).

Attempts to Affect Recognition with Feedback

Inspired by the idea that even the most simple recognition memory task involves subtly nuanced and sophisticated generation and monitoring processes, in the late 1990s our lab began studying the possibility that recognition discrimination could be fine-tuned via accuracy feedback at test. In our most basic procedure, undergraduate subjects studied a long list of words presented one at a time on a computer screen and then later were shown a randomly ordered mix of studied and non-studied words presented one at a time for recognition judgments. In most of these studies, subjects responded on a 6-point scale from "sure not studied" to "sure studied." A randomly selected half of the subjects were given accuracy feedback after each test response (e.g., if the item was new and they rated it on the "new end" of the scale they were told "Correct! That item was NOT on the study list"). Note that in this procedure there are no repeated study test cycles (unlike, say, Jennings & Jacoby's [2003] memory training procedure, in which effects of feedback might be mediated by changes in how items are studied). In our basic procedure there is just one study list followed by one test list in which each test item is presented once. Our hypothesis was that subjects could use feedback to fine-tune the way they engaged with probe words and/or the way they evaluated their memorial responses to probe words, leading to a gradual increase in accuracy over the course of the test.

A number of researchers have used feedback procedures in the context of various memory tasks, but only a few have tested the hypothesis that feedback can enhance recognition discrimination without item repetition or repeated study-test cycles (see Kantner & Lindsay, 2009, for review). Titus (1973) used CVCs and test lists in which only 20% (15 of 75) of the items were old. He found that subjects who received accuracy feedback after each test response adopted a more conservative response criterion but were not significantly more accurate than control subjects. More recently, Verde and Rotello (2007) and Han and Dobbins (2008) reported null effects of feedback on recognition sensitivity, but their studies were primarily oriented toward testing hypotheses regarding effects of feedback on response criterion (which they did obtain).

We set out to explore conditions designed to maximize the likelihood of observing effects of feedback on recognition sensitivity. Our initial strategy involved using multiple words from each of several semantic categories (e.g., birds, minerals), with each category represented by studied and nonstudied items, with the aim of making old/new discrimination rather difficult despite relatively deep processing at study. We predicted that subjects would use feedback to adjust the ways in which they engaged with probe words (thereby improving the diagnosticity of the "echo" generated from memory) and/or to improve their monitoring/evaluation processes (thereby better discriminating between their own internal responses that were predictive of oldness vs. newness). Thus we expected that over the

course of the test subjects who received feedback would gradually come to be more accurate than control subjects in their old/new judgments.

There was not a whiff of a feedback effect in our initial experiment, which proved to be the first of an extensive collection of null effects (see Kantner & Lindsay, 2009, for a few select specimens). We tried various numbers of items, orienting tasks, delays, ways of giving the feedback, etc. Nothing, neither in terms of accuracy nor confidence. Feedback combined with proportion-old manipulations affected response criterion, presumably because feedback acts as a proportion-old instruction, but we found no other effects of feedback on recognition performance.

We then started employing less standard recognition situations. In one set of studies, we presented two study lists but instructed subjects they were to recognize only items from one of those lists even though the test list included some items from the other list, which subjects were to reject (as in Jacoby's opposition procedure; e.g., Jacoby, Woloshyn, & Kelley, 1989). Performance on this test was unaffected by feedback. We also conducted studies (inspired by Higham & Brooks, 1997) in which, for some subjects, old words were names of large objects and new words were names of small objects (or vice versa for other subjects); compared to mixed list, this confound led to slightly better recognition performance, but that effect was not amplified by feedback. In another set of studies, we used a variant of the Deese (1959) procedure in which subjects study the second through fifth most often generated exemplars of each of a large number of categories (e.g., precious metals: silver, bronze, copper, platinum) and then take a test that includes, as the critical items, the most often generated exemplar of studied (e.g., gold) and non-studied (e.g., robin) categories (Seamon, Luo, Schlegel, Greene, & Goldenberg, 2000; Smith, Ward, Tindell, Sifonis, & Wilkenfeld, 2000). As other researchers have shown, false alarm rates were high for the critical items from studied categories. Subjects who received feedback were just as likely to false alarm to critical items as were other subjects, even though every time they endorsed such an item they were told that they were wrong.

Over the years, while this recalcitrant project simmered on the back burner, we sought the advice of a number of renowned memory researchers. Some were incredulous that we would entertain the idea that recognition memory discrimination could be tuned by feedback at test without item repetition or repeated study/test cycles. From their perspective, recognition memory is so encapsulated and automatized (at least in adults) that it is impervious to such manipulations. Maybe they are right – to anticipate, to this day we have not firmly nailed down a reliable effect of feedback (without item repetition or repeated study/test cycles) on recognition discrimination (but read on for some tantalizing hints). Other scholars offered suggestions for ways of "shining the light" (to quote social psychologist Lee Ross) so as to reveal such an effect. Chris Herzog, for example, speculated that subjects might benefit if instead of merely giving them accuracy feedback we also told them their response latency for each old/new response (on the ground that, empirically, quick responses are more often accurate than slow ones [e.g. Koppell, 1977]). We tested this idea a couple of times with no hint of an effect. Asher Koriat proposed that if we mixed markedly high and low frequency items, subjects receiving feedback might better tune in on the fact that recognition discrimination is more difficult for high frequency items (and hence at least respond with lower confidence on such items). Once again, we observed no such effect. Mike Masson suggested that we might obtain feedback effects if we tested with 2-alternative forced choice recognition rather than yes/no recognition, but here too we found null effects of feedback.

Bruce Whittlesea speculated that subjects may already be optimized in their ability to recognize words; years and years of experience with these materials may have tuned the system as much as it could be tuned. Thus, Bruce proposed that we might observe feedback effects if we used stimulus materials that have a rich and complex structure but are unfamiliar to our subjects. The idea here is that such stimuli will give rise to ambiguous mental productions at test; being strange and complicated, the test items will be somewhat difficult to process and subjects' internal responses to them rather confusing (e.g., "Is this a stimulus I heard before, or is it just that I sort of like it?"). Bruce's idea set us off on a journey that we recount in the remainder of this chapter. The first stop is Korea.

Korean Melodies

To the untutored Western ear, traditional Korean music is passing strange. The instrumentation, scales, rhythms, and structures all depart from those of European music (visit http://203.252.231.26/

for samples). There are several major genres, some being orchestral in quality, others sounding folksy, and yet others reminiscent of free jazz. The music is evocative and challenging; listening to it is not a neutral experience. It struck us as the perfect medium with which to explore Bruce's idea that effects of feedback on recognition sensitivity might be observed with complexly structured yet deeply novel materials.

In an initial foray into this domain, we created 48 10-second mp3 files, each sampled from a different piece of Korean music (collected from the internet and the UVic library). We randomly divided them into two sets of 24 files each. Subjects were asked to listen to each of the melodies in one of the sets, with a few seconds between each one, under instructions to remember the music for a subsequent test. The music was played over the computer's built-in speaker. Then, after a 1-minute filled delay and the test instructions, each of the 48 melodies was played in a fixed, randomly intermixed order, and subjects made yes/no recognition memory judgments, giving their responses aloud. In this and all of the experiments we have conducted on the effects of feedback on recognition memory, the subjects were University of Victoria undergraduates who participated for optional bonus points in psychology courses or for a nominal payment. In each of the studies reported here, a quasi-randomly selected half of the subjects received accuracy feedback after each recognition judgment.

Mean recognition accuracy (d') in the feedback and control conditions is shown in the far left pair of bars in the upper panel of Figure 1A; mean response bias (c) estimates are presented in the lower panel (negative scores indicate a conservative bias).[1] Error bars represent the standard error of the mean. Breakthrough! Just as Whittlesea had predicted, we obtained a statistically significant effect of feedback on the accuracy with which Korean melodies were recognized (with no effect of feedback on response criterion, which was essentially neutral).

Given the large number of null effects that had preceded this finding, we had little confidence in it, suspecting that it would prove to be a Type I error. We conducted a conceptual replication with a larger *N* and a number of minor improvements (e.g., stimulus presentation and feedback administration were more carefully controlled via E-Prime, responses were entered into the computer instead of being spoken aloud, and assignment of items to studied vs. new status was randomized anew for each subject). Experiment 2 was conducted in a different semester and by a different experimenter than Experiment 1. We were delighted when the effect of feedback on accuracy replicated, as shown in the second pair of bars in the upper panel of Figure 1. This time there was also a significant tendency for subjects who received feedback to be more conservative on the test, but the crucial finding from our perspective was that, unlike all of our prior studies with words, feedback significantly improved recognition discrimination for Korean melodies, just as Bruce had predicted.

Maybe we should have written up this pair of studies and submitted them for publication, but we instead launched several lines of investigation designed to explore for feedback effects in recognition memory for other rich, temporally extended, and/or emotionally evocative stimuli (some of which are described below). For a variety of reasons we subsequently returned to our Korean melody (KM) materials, conducting three additional studies with the assistance of student Danette Dawkin. The first two KM studies, reported above, had used yes/no response alternatives at test, whereas most of our other feedback studies had used a 6-point rating scale. Experiment 3 was essentially the same as Experiment 2, except that we tested 40 subjects and a randomly selected half of them responded on a 6-point rating scale (whereas the remaining subjects made yes/no responses). Also, relative to Experiments 1 and 2, KM3 was conducted in a different semester and by a different experimenter and subjects in KM3 used headphones to listen to the audio clips at study and test. That same experimenter in that same semester then conducted Experiment 4, a replication of KM3 including only the yes/no response conditions and eliminating the use of headphones on the off chance that they made a difference.

[1] Hit and false alarm rates for all experiments are presented in Appendix A. We tried various measures of sensitivity and criterion. For the experiments in which subjects responded on 6-point scale, we also analyzed the data in terms of mean scale response. None of those analyses suggested any different story than did d' and c, so for simplicity we report only those two measures here. Email either author for the mean scale responses or the raw data.

As shown in the three pairs of bars on the right of the upper panel of Figure 1, each of these three comparisons yielded a nonsignificant tendency toward an anti-feedback effect. That is, we failed to replicate the beneficial effect of feedback on accuracy obtained in the first two KM experiments, regardless of test format. The lower panel of Figure 1 shows that, as in Experiment 2, both test-form conditions of Experiments 3 and 4 yielded a significant effect of feedback on response criterion: Subjects who received feedback after each test response were more reluctant to judge melodies old than were subjects who did not receive feedback. We defer further discussion of the Korean melodies findings pending report of two other parallel lines of investigation.

Masterwork Paintings

A great painting, like a provocative piece of music, can be emotionally moving in complex ways. One may, for example, be enthralled by the technical skill of a painting yet appalled by its subject matter (e.g., Rembrandt's The Anatomy Lesson of Dr. Tulp) or one may love the colours and textures of a painting but not really "get" the forms (e.g., some might feel that way about Rothko's ephemeral paintings). The complexity and evocativeness of great paintings led us to speculate (along the lines of Bruce Whittlesea's suggestion) that we might obtain effects of feedback on recognition of paintings.

Jeffrey Toth (of the University of North Carolina Wilmington) has assembled a large set of standardized digital scans of great paintings for use in a different sort of memory training program, and he kindly allowed us to use images from that set. For our initial study with these materials, we selected 102 paintings. We avoided super-famous works such as the Mona Lisa, and sampled a wide range of styles, artists, and subject matter. The experiment was conducted by student Kyle Mathewson. In the study phase, 54 paintings were presented one at a time for 1 sec each with instructions to "study each painting carefully and completely." Immediately after the study phase, the test instructions were given, and then the studied paintings (less 6 primacy and recency buffers) and an equal number of nonstudied paintings were presented one at a time in a novel randomized order and subjects used a 6-point scale to make confidence-weighted recognition memory judgments (sure new to sure old). As shown in the leftmost pair of bars in the upper panel of Figure 2, feedback had no effect on recognition sensitivity. But, as in the latter four Korean Melodies experiments, subjects who received feedback displayed a significantly more conservative response criterion than did control subjects.

Recognition accuracy was quite high in Paintings Experiment 1, and we speculated that this may have made it difficult to detect a beneficial effect of feedback on sensitivity. We therefore devised a new version of the experiment, in which all of the paintings were portraits. Relative to Paintings Experiment 1, Paintings Experiment 2 was conducted in a different semester by a different experimenter (student Elaine Blight), and included a larger number of study and test items (75 each, plus 10 primacy and recency buffers). As can be seen in Figure 2, the shift to using only portraits and the increase in the number of items apparently had the desired effect of substantially lowering response accuracy, but there was no effect of feedback on accuracy. Moreover, the large effect of feedback on response criterion observed in Paintings Experiment 1 (and in most of the KM experiments) was eliminated. Elaine Blight then conducted an "exact" replication of Paintings Experiment 1. This Paintings Experiment 3 yielded no effect of feedback on accuracy or response criterion. Then Elaine conducted yet another study, which was identical to Paintings Experiment 2 (i.e., with only portraits) except that a deep orienting task was used during the study phase. Specifically, subjects in Paintings Experiment 4 were asked to rate whether the individual depicted in each portrait looked above or below average in friendliness, with an eye to increasing recognition sensitivity with the homogeneous portrait set. Again, neither sensitivity nor criterion effects of feedback were observed.

In a subsequent semester, new research assistants Brian Buchan and Alison Wegner conducted yet another study using the paintings materials in which we attempted to manipulate how motivated subjects were to perform well on the recognition test. Our speculation was that the experimenter who conducted the first Paintings study may have inspired subjects to try harder on the test, and that this may have led to the pronounced effect of feedback on response criterion. The findings of four of the KM studies and Paintings Experiment 1 suggest that when subjects receive feedback they try especially hard to avoid false alarms (at the cost of increased misses); we speculated that this feedback-driven conservative shift would be particularly strong if subjects were highly motivated to do

well. But Brian and Alison found no effect of feedback on recognition test responses in Paintings Experiment 5, regardless of motivation condition.

In Paintings Experiment 5, we included a manipulation check at the end of the procedure, asking subjects to rate how motivated they had been to do well on the recognition test. That measure indicated that our attempt to manipulate motivation was ineffective (on a 7-point scale with 7 = highly motivated, means of 5.7 and 5.8 in the high- and low-motivation conditions, respectively). Self-reported motivation was also equivalent in the feedback (M = 5.8) and control (M = 5.7) conditions. Response bias was marginally more conservative for participants receiving feedback, but only in the low motivation condition ($p < .07$), and no other remotely significant bias or sensitivity effects were observed. There was a modest and significant correlation between self-reported motivation and d' ($r = .25$, $p < .01$), such that individuals who reported higher motivation tended to be more accurate at discriminating studied from new paintings. But contrary to our speculation about motivation playing a role in the conservative bias, there was no significant correlation between self-reported motivation and c and the miniscule relationship we did observe was in the wrong direction ($r = .10$). At this point we downed brushes.

Poetry

T. S. Eliot wrote, "Genuine poetry can communicate before it is understood." Still inspired by Bruce Whittlesea's idea that effects of feedback on recognition might arise with complex, quixotic materials, we assembled a sample of lines of poetry by Rainer Maria Rilke. For example:

> The walls, with their ancient portraits, glide
>
> away from us, cautiously, as though
>
> they weren't supposed to hear what we are saying.
>
> From "Before Summer Rain"

Our intuition was that intro psych students might find it challenging to read dozens of snippets of Rilke by themselves. We therefore created +/- 10-s audio clips of the lines being read aloud somewhat poetically—not overly so, but with a degree of feeling. During the study phase, each of 54 clips, randomly selected anew for each subject from a set of 96 clips, was played aloud, with a 500-ms pause between clips. Test instructions were presented immediately after the study list, and then each of the studied clips (less 6 primacy and recency buffers) was again played aloud, randomly intermixed with an equal number of nonstudied clips. A total of 47 subjects responded to each clip on a 6-point scale from sure new to sure old. Half of the subjects received accuracy feedback following each response.

The mean recognition accuracy scores are shown in the first pair of bars in the upper panel of Figure 3. As in Korean Melodies Experiments 1 and 2, subjects who received feedback as they took the test had significantly higher recognition accuracy than did control subjects. As shown in the lower panel of Figure 3, we also once again found that subjects who received feedback were significantly more conservative than control subjects.

In two subsequent studies, identical to one another, the same study and test poetry snippets as in Poetry Experiment 1 were presented as text on the computer screen. At study subjects were instructed to read each poetry snippet silently to themselves and to study it for a later memory test. Likewise at test subjects read each snippet silently and responded with the 1-6 rating scale. There were 31 and 34 subjects in Poetry 2 and Poetry 3, respectively. As shown in Figure 3, this text-only version of the Poetry procedure yielded mixed effects of feedback. Feedback significantly impaired recognition sensitivity in Poetry 2 but not in Poetry 3. Also, as in many of the studies we have reported, subjects who received feedback were directionally more conservative in their recognition judgments, but in neither of these experiments did that tendency attain statistical significance.

As mentioned above, in Poetry Experiment 1 the Rilke audio clips had been read "poetically." We wondered if this emotive expressiveness may have played a role in giving rise to the beneficial effect of feedback on recognition accuracy in that experiment. To explore this possibility, we conducted a follow-up experiment in which half of the subjects heard the same audio clips as in Experiment 1 whereas other subjects heard the same poetry snippets rendered by a flat "robot" voice (using TextAloud software). Experiment 4 Human

directionally replicated Experiment 1, but the tendency toward on effect of feedback on d' did not approach significance (p = .19) and even this estimate exaggerates the tendency toward a feedback effect on sensitivity because d' values were skewed due to very low false alarm rates (see Appendix): In terms of hits minus false alarms, there was no difference at all between the conditions. Thus Poetry Experiment 4: Human Voices (which was essentially an exact replication of Poetry Experiment 1) failed to replicate the effect of feedback on sensitivity. In contrast, the feedback-based conservative-shift effect obtained in Poetry Experiment 1 was even larger in Poetry Experiment 4: Human Voices. In Poetry Experiment 4: Robot, there was no hint of an effect of feedback on recognition accuracy, but the conservative-bias effect was comparable to that observed in Poetry Experiment 1.

Summary Sans Conclusions

In this chapter we have taken what we suspect is an unusual tack by reporting every experiment we have conducted along these particular sub-lines of research.[2] That is, aside from a few small-*N* pilot studies, this is an exhaustive catalog of our studies of the effects of feedback on recognition of Korean melodies, poetry, and paintings. By Whittlesean standards the number of studies is modest, but it is sufficient to create a rather complex picture. Indeed, the pattern of results is maddeningly inconsistent. If you are holding your breath waiting for us to provide a coherent theoretical account for this patchwork of null and significant effects, you will soon turn blue. Nonetheless we think our findings make some important points.

Can feedback at test enhance recognition sensitivity? Is such an enhancement effect more likely to be observed with stimuli that are rich, complex, and poorly understood, as Bruce Whittlesea surmised? Maybe a little. Against the backdrop of the forest of null effects obtained in our studies with familiar words as stimuli (Kantner & Lindsay, 2009), the three or four isolated poplars of significant effects with Korean melodies and poetry clips stand out in stark contrast. But even if these effects are real, rather than Type I errors, the effect is far from robust. Indeed, a mega-analysis of the raw data collapsed across the 16 experiments revealed virtually identical mean d' values for subjects who did and did not receive feedback (M = 1.469 and 1.468, $F(1, 536) < .001$, p = .99, partial eta squared < .001). Also, in the Poetry studies in which there was at least some indication of better discrimination by subjects who received feedback, there was no clear indication that discrimination improved gradually over trials (we judged that the KM studies had too few trials to afford such analyses). Of course, it is always possible that larger and more consistent benefits of feedback on sensitivity would be observed under other conditions, but such conditions have thus far eluded us despite an extended search.[3]

In contrast to the rather murky findings in the analyses of recognition sensitivity, the data on response bias are strikingly consistent in two regards. First, subjects tended to be conservative, preferring misses to false alarms (or, to put it the other way, placing a higher premium on correct rejections than on hits). Of the 32 comparisons reported here, response criterion was at least directionally conservative in 26, and significantly so in 20. Collapsing across experiments and conditions, the tendency for c to be less than zero was significant and substantial (M = -0.232, $t(538)$ = -15.7, p < .001). In all of these experiments, half of the test items had been studied, and there was no explicit incentive to value correct rejections above hits. Although we have not systematically assessed the matter, a casual review of our studies with familiar words indicates no such tendency; lacking relevant manipulations (e.g., of the base rate of old vs. new items), in our studies with words response criterion has been

[2] We also conducted an experiment with Chinese characters (with the assistance of students Jeffrey Sun, Ben Shiner, and Danika Overmars), a study with pseudowords (with Ben Shiner), a study with "one-liners" (e.g., "A drunk man's words are a sober man's thoughts" or "43% of statistics are worthless") (with students Brian Buchan, Kyle Mathewson, and Nicky Schnare), and a study with photos of faces (with Nicky Schnare). None of these studies yielded an effect of feedback on recognition sensitivity or response bias (although there was a nonsignificant tendency toward a feedback-based conservative-shift effect in the study with faces).

[3] Lane, Roussel, Villa, and Morita (2007) reported a fascinating experiment using feedback in the eyewitness misinformation domain. Subjects viewed a slide series depicting an event, were exposed to misleading suggestions regarding details in that event and either did or did not receive feedback on the first part of a test on which subjects were to identify the source (slides, narrative, both, or neither) of details from each of those sources. Accurate feedback selectively lowered false attributions of suggested details to the slides, without evidence of a general change in response criterion.

near zero. We plan a future experiment designed to test for a hypothesized materials-based bias shift in recognition memory for words versus paintings and other such stimuli. Note that others have found that more memorable stimuli often evoke more conservative response bias (e.g., Singer, 2009; Stretch & Wixted, 1998)

The second clear finding pertaining to response criterion was that receiving feedback substantially increased subjects' conservative bias. This feedback-based conservative-shift effect was directionally present in 14 of the 16 experiments, and statistically significant in 7 (and "marginal" in one more) of those comparisons. Collapsing across experiments, the difference in c between the feedback (M = -.32) and control (M = -.15) conditions was significant, $F(1, 536) = 44.17$, $p < .001$, partial eta squared = .08). Here again, this finding contrasts with the results from our studies with familiar words, in which response bias did not consistently differ between subjects who did and did not receive feedback.

Why should Korean melodies, poetry clips, and paintings give rise to a conservative response criterion? Perhaps the evident richness and distinctiveness of the stimuli leads subjects to believe that they should have good memory for them, and hence should be able to recollect a lot about studied items when probed. That is, subjects' expectations about what they should be able to remember about a test item might be exaggerated, leading them to reject a high proportion of test items. Or perhaps subjects believe that to claim to have experienced such a stimulus when one hadn't would be a particularly egregious kind of error. In pondering this issue, we were struck by the parallel to classroom situations in which students are loath to volunteer an answer unless they are very confident of it, perhaps following the proverb "Even a fool, when he holdeth his peace, is counted wise, and he that shutteth his lips is esteemed a man of understanding" (Proverbs 17:28, King James Bible). Whatever the explanation, to the best of our knowledge no previous publication has reported evidence of a materials-based conservative-shift effect on recognition memory response criterion.

Why should feedback amplify the tendency toward conservative bias on tests of recognition memory for these sorts of stimuli? At first blush, one might have thought that feedback would correct, rather than exaggerate, response bias. After all, feedback informed subjects that they were more often erring by rejecting studied items than by false-alarming to new items. Yet the effect of receiving feedback was to increase that conservative bias. This is arguably consistent with the speculation that subjects were particularly motivated to avoid false alarms.

Whittlesea (2002, p. 112) wrote provocatively about response bias in recognition memory, arguing that although signal detection models often fit data beautifully they do not necessarily adequately capture the underlying processes that modulate participants' attitudes toward their own responses to test probes. More specifically, he proposed that criteria may shift fluidly from trial to trial, "on the fly," influenced by interactions between characteristics of the item, the person, and the context. Whittlesea's arguments here were on the cutting edge, as numerous researchers have addressed the issue of criterion flexibility in a variety of contexts in recent years (e.g., Curran, DeBuse, & Leynes, 2007; Dobbins & Kroll, 2005; Heit, Brockdorff, & Lamberts, 2003; Gallo, Roediger, & McDermott, 2001; Hockley & Niewiadomski, 2007; McCabe & Balota, 2007; Miller & Wolford, 1999; Rhodes & Jacoby, 2007; for recent reviews of response bias in recognition memory, see Hockley, this volume; Rotello & Macmillan, 2008).

Caveat

The appalling lack of consistency across our studies in the effect of feedback on recognition discrimination provides an important albeit not novel lesson. Consider, for example, the first three of our Korean melodies studies. With respect to recognition sensitivity, KM Experiment 2 nicely replicated KM Experiment 1, but KM Experiment 3: Scale Response yielded an anti-feedback effect. The only planned, systematic difference between the latter two of these experiments was in the test response alternatives. If we (a) indulged in across-experiment comparisons and (b) did not also include in KM Experiment 3 a yes/no condition that produced exactly the same anti-feedback effect as did the scale response condition, we would have been mightily tempted to tell a just-so story in which responding on a scale reverses the beneficial effects of feedback. But KM Experiment 3: Yes/No was a quite close replication of KM Experiments 1 and 2, and we can only speculate as to what caused the reversal in the data pattern. Perhaps mere measurement error in one or both experiments. Similarly, whereas the first of our studies with paintings yielded the most dramatic effect of feedback on conservative bias, a

near-exact replication of that study produced no hint of such an effect. Here again, we can only speculate as to why.

The scary thing about this is that it would be very easy for researchers collecting data in a noisy domain such as this to fall into error, knowingly or unknowingly. We could, for example, have made a very compelling case for the reality of effects of feedback on d' by omitting mention of the experiments that "didn't work." Or we could have gone even further by reporting several studies with an effect of feedback on d' and one or two contrasting null-effect studies, interpreting the latter as setting boundary conditions on the effect of feedback on sensitivity. For example, by selecting from these studies one could tell a nice story to the effect that feedback enhances recognition sensitivity with temporally extended auditory stimuli such as Korean melodies and poetry, but not with static stimuli such as paintings. There may even be something to that story, but the fact that we several times failed to observe an effect of feedback on Korean melodies and poetry undermines confidence in it. At this point, despite substantial effort we don't know what caused the effect to sometimes emerge and other times not.

One implication is that researchers are well advised to determine empirically whether or not the effects they find are replicable, ideally with more than one replication. The fact that a finding is statistically significant does not mean that it is replicable (Miller, 2009). It is not unusual for cognitive psychologists to tweak a procedure until it "works," which is fine as long as one goes on to determine whether or not it works *consistently*. A related implication is that we must be cautious is making across-experiment comparisons. That point is almost too elementary to warrant mention, but the field is rife with insufficiently qualified across-experiment comparisons.

Last Word

Pursuing this line of research has been an intellectual and emotional rollercoaster ride. The significant effect of feedback on recognition sensitivity in the first two Korean melody studies inspired by Bruce Whittlesea's suggestion produced a heady high that lingers on in memory. That this effect again emerged in the first of our poetry studies and (marginally) in the replication of that study (i.e., in Poetry Experiment 4: Human Voices) continues to intrigue and encourage us. But our replication failures (which occurred despite our use of highly standardized, computer-run procedures) have at times plunged us to the depths, and at this point we are still far short of understanding the effects of feedback on recognition discrimination.[4]

The fickleness of our discrimination effects was offset by our consistent findings regarding response bias. The suggestion of a materials-based conservative-shift effect (which to date rests solely on across-experiment comparisons!) was an unlooked-for bonus of this line of research. Likewise, the quite strong evidence of a feedback-based conservative-shift effect that emerged in these studies is a serendipitous discovery that may contribute to developments in theorizing about response criterion in recognition memory (Hockley, this volume; Rotello & MacMillan, 2008).

Research sometimes teaches us how little we know. Can feedback enhance recognition discrimination? Is such an effect more likely with certain kinds of materials than others, and if so what are the determining factors? What factors accounted for the inconsistency of the sensitivity effects across experiments? Why is bias more conservative with rich, complex stimuli such as music or poetry than with words? Why does feedback encourage a conservative response bias with such materials? Any ideas, Bruce?

[4] Zolton Dienes, in a review of a draft of this chapter, raised the intriguing possibility that fluke variations in item-order effects might have contributed to the inconsistency of our results.

References

Cohen, R. L. (1989). Memory for action events: The power of enactment. *Educational Psychology Review, 1*, 57-80.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 2-17.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1186-1198.

Gallo, D. A., Roediger, H. L. I.,II, & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8*, 579-586.

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

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10*, 718-723.

Hockley, W. E. (this volume). Criterion changes: How flexible are recognition decision processes? Chapter in P. Higham and J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea*. Houndmills, UK: Palgrave Macmillan.

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35*, 679-688.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306-340.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391-422). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118*, 126-135.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General, 118*, 115-125.

Jennings, J. M., & Jacoby, L. L. (2003). Improving memory in older adults: Training recollection. *Neuropsychological Rehabilitation, 13*, 417-440.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3-28.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology, 94*, 37-64.

Kantner, J., & Lindsay, D. S. (2009). *Can corrective feedback improve recognition memory?* Manuscript under review for publication.

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory, vol. 41* (pp. 293-320). San Diego, CA US: Academic Press.

Koppell, S. (1977). Decision latencies in recognition memory: A signal detection theory analysis. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 445-457.

Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. K. (2007). Features and feedback: Enhancing metamnemonic knowledge at retrieval reduces source-monitoring errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 1131-1142.

Lindsay, D. S. (2008). Source monitoring. In H. L. Roediger III (Ed.), *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference,4 vols.* (J. Byrne, editor) (pp. 325-348). Oxford: Elsevier.

McCabe, D. P., & Balota, D. A. (2007). Context effects on remembering and knowing: The expectancy heuristic. *Journal*

of Experimental Psychology: Learning, Memory, and Cognition, 33, 536-549.
Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. Psychological Review, 106(2), 398-405.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*, 519-533.
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.
Roberts, K. P., & Blades, M. (2000). *Children's source monitoring*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.
Rotello, C. M., & Macmillan, N. A. (2008). Response bias in recognition memory. In A. S. Benjamin, B. H. Ross, A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use* (pp. 61-94). San Diego, CA US: Elsevier Academic Press.
Seamon, J. G., Luo, C. R., Schlegel, S. E., Greene, S. E., & Goldenberg, A. B. (2000). False memory for categorized pictures and words: The category associates procedure for studying memory errors in children and adults. *Journal of Memory and Language, 42*, 120-146.
Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*, 976-984.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592-604.
Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition, 28*, 386-395.
Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1379-1396.
Titus, T. G. (1973). Continuous feedback in recognition memory. *Perceptual and Motor Skills, 37*, 771-776.
Unkelbach, C. (2006). The learned interpretation of cognitive fluency. *Psychological Science, 17*, 339-345.
Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 219-230.
Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*, 254-262.
Whittlesea, B. W. A. (2002a). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131*, 96-115.
Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica, 98*, 141-165.
Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 547-565.
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.

Author Note

Please address correspondence to Steve Lindsay, Psychology, UVic, P.O. Box 3050 STN CSC, Victoria, BC, V8T 2M5, Canada. E-mail slindsay@uvic.ca. Phone 1-250-721-8593. Fax 1-250721-8929. Or to jkantner@uvic.ca.

We thank the many students who helped prepare stimulus materials and test research participants. We also thank Bill Hockley and Zolton Dienes for insightful comments on a draft of this chapter. And thanks to Bruce Whittlesea for being such an inspiration.

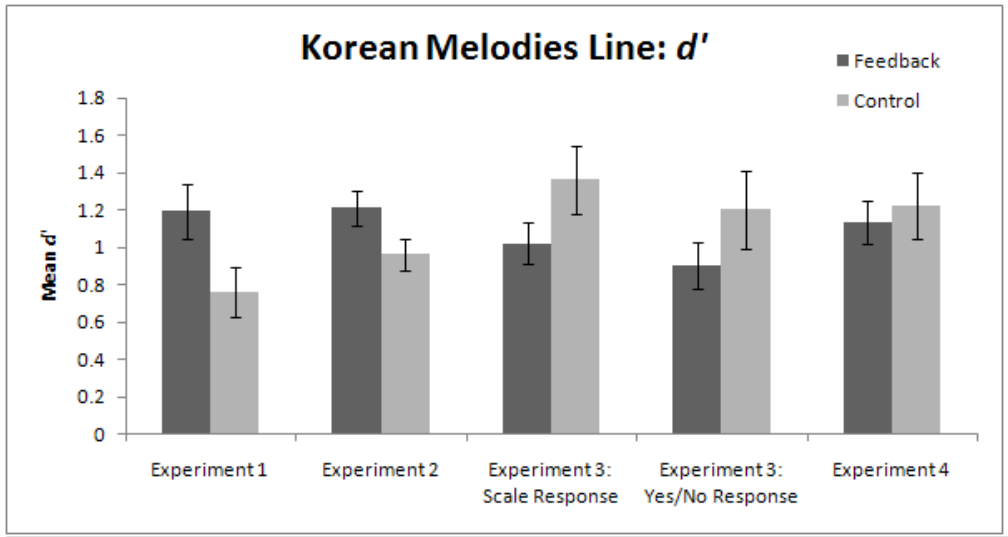
Figure Captions

Figure 1. Upper panel: Mean values of d' (and standard errors of the mean) in the feedback and control conditions of the Korean Melodies experiments. Lower panel: Mean values of c (negative is conservative).

Figure 2. Upper panel: Mean values of d' (and standard errors of the mean) in the feedback and control conditions of the Paintings experiments. Lower panel: Mean values of c (negative is conservative).

Figure 3. Upper panel: Mean values of d' (and standard errors of the mean) in the feedback and control conditions of the Poetry experiments. Lower panel: Mean values of c (negative is conservative).

Figure 1


Korean Melodies Line: d'
Feedback
Control
Mean d'
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Experiment 1
Experiment 2
Experiment 3: Scale Response
Experiment 3: Yes/No Response
Experiment 4

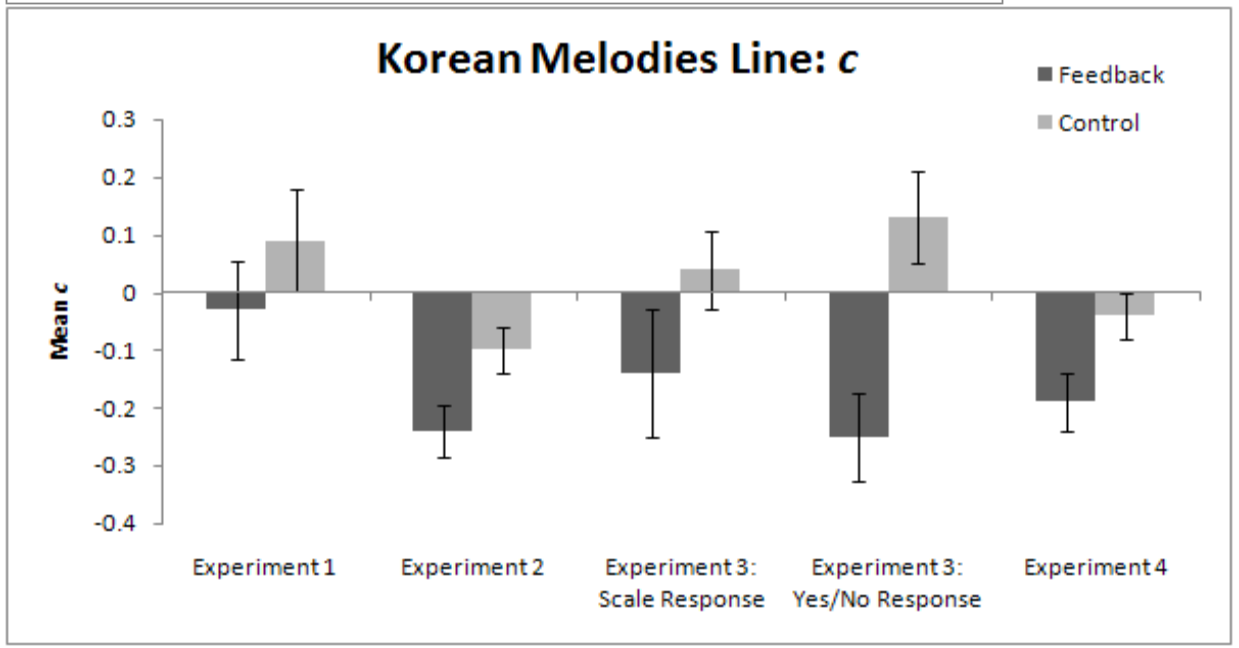
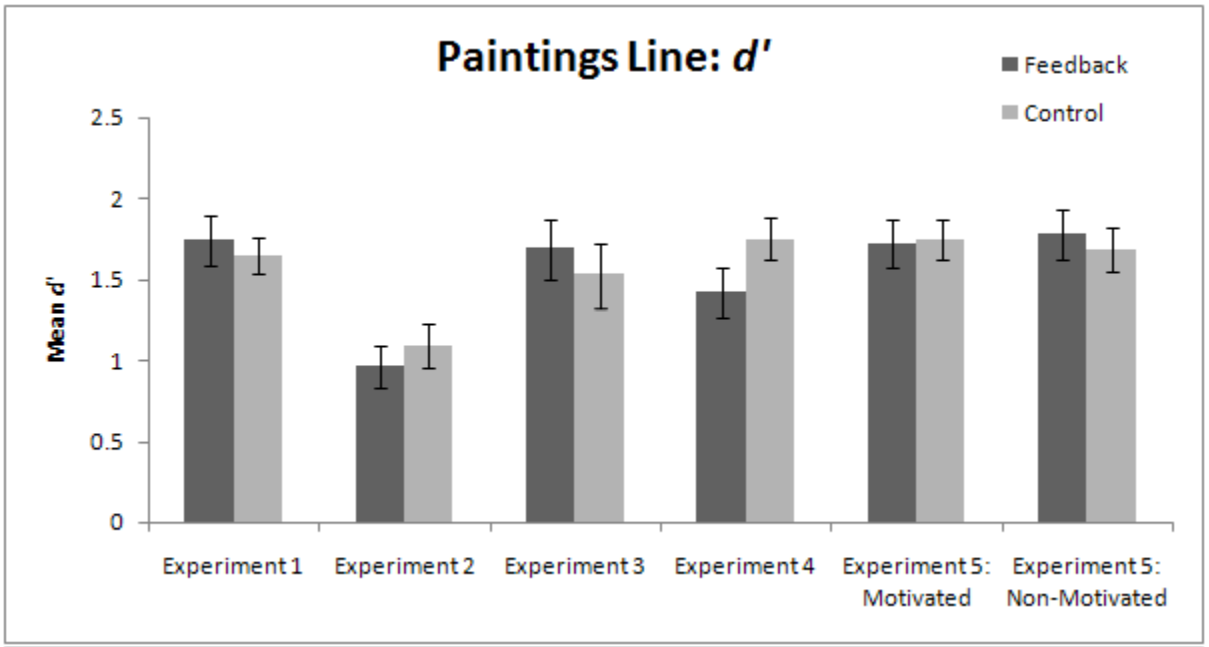

Korean Melodies Line: c
Feedback
Control
Mean c
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Experiment 1
Experiment 2
Experiment 3: Scale Response
Experiment 3: Yes/No Response
Experiment 4

Figure 2


Paintings Line: d'
Feedback
Control
Mean d'
2.5
2
1.5
1
0.5
0
Experiment 1
Experiment 2
Experiment 3
Experiment 4
Experiment 5: Motivated
Experiment 5: Non-Motivated

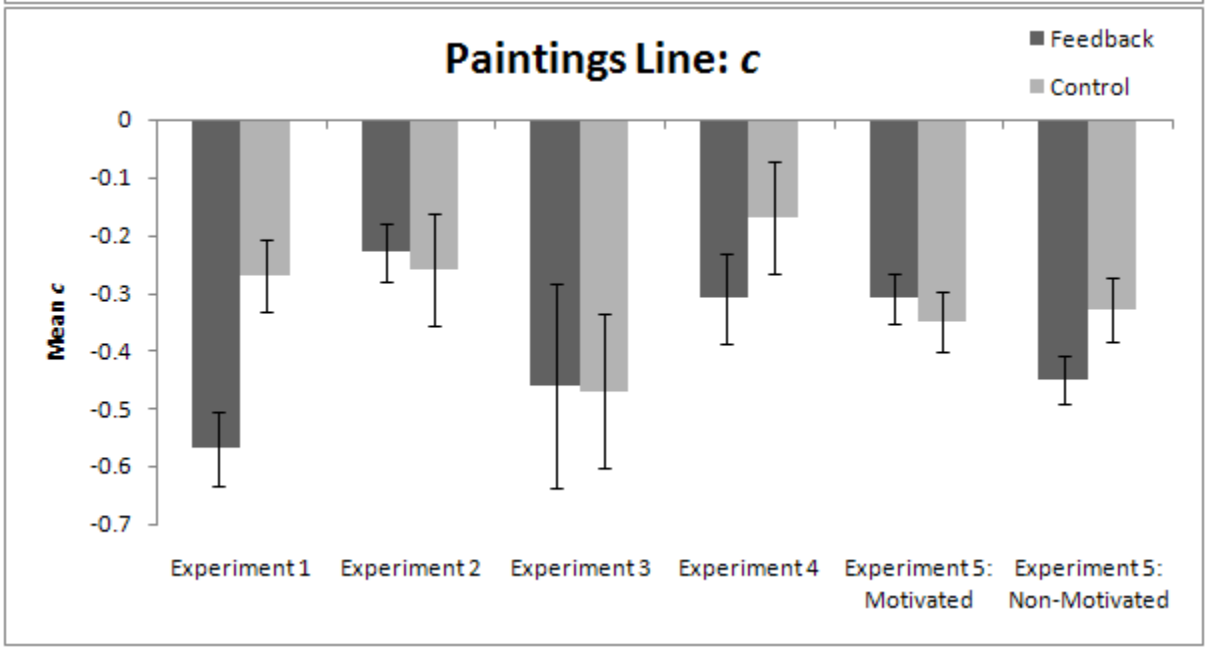
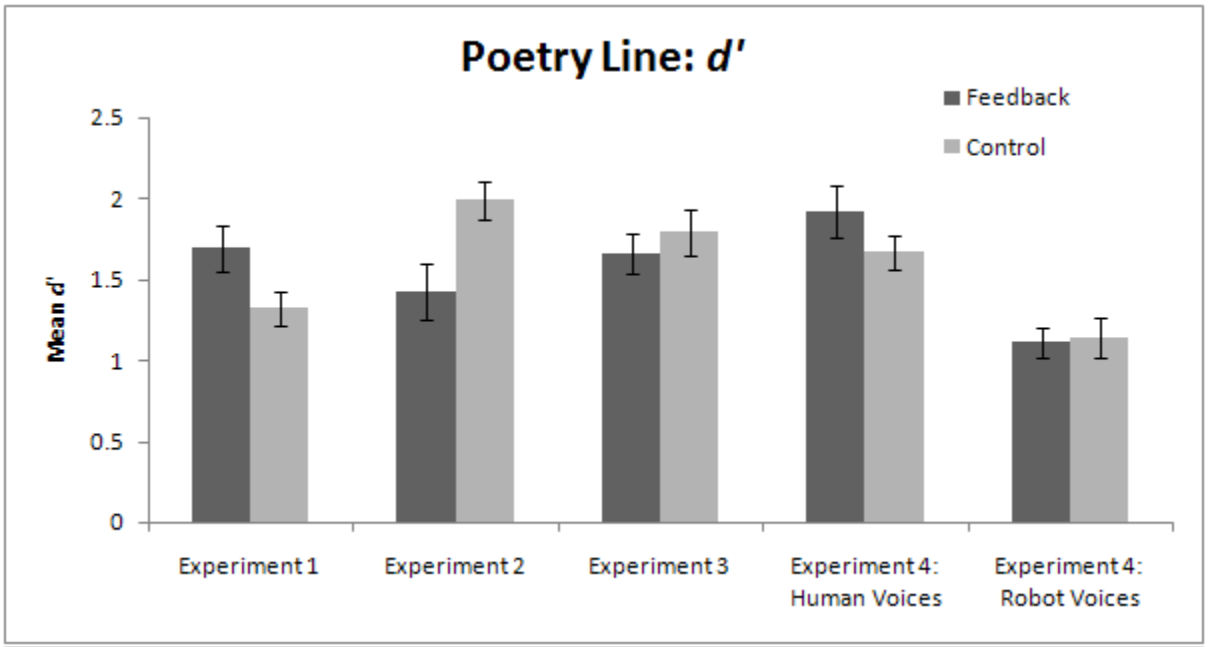
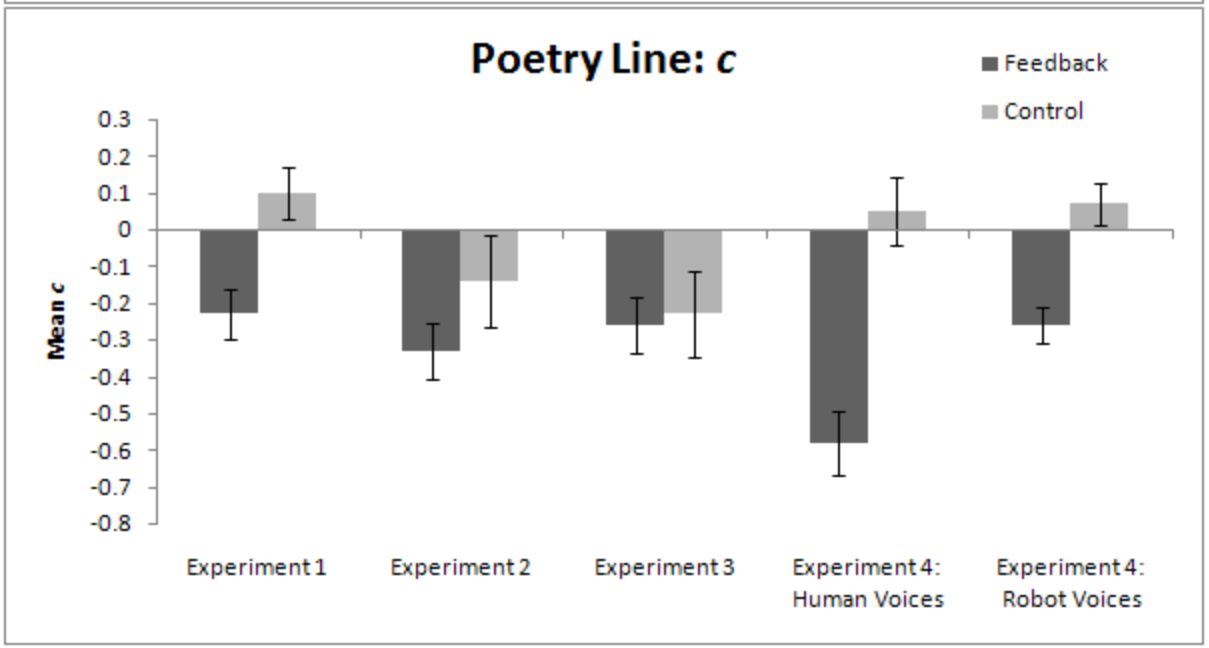

Paintings Line: c
Feedback
Control
Mean c
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
Experiment 1
Experiment 2
Experiment 3
Experiment 4
Experiment 5: Motivated
Experiment 5: Non-Motivated

Figure 3


Poetry Line: d'
Feedback
Control
Mean d'
2.5
2
1.5
1
0.5
0
Experiment 1
Experiment 2
Experiment 3
Experiment 4: Human Voices
Experiment 4: Robot Voices


Poetry Line: c
Feedback
Control
Mean c
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
Experiment 1
Experiment 2
Experiment 3
Experiment 4: Human Voices
Experiment 4: Robot Voices

Appendix

Experiment	H		FA	
	Feedback	Control	Feedback	Control
Korean Melodies Line				
Experiment 1	.708 (.035)	.675 (.030)	.278 (.039)	.396 (.041)
Experiment 2	.638 (.015)	.646 (.017)	.217 (.022)	.292 (.022)
Experiment 3: Scale	.640 (.028)	.755 (.032)	.280 (.044)	.275 (.039)
Experiment 3: Yes/No	.575 (.038)	.740 (.044)	.250 (.028)	.325 (.027)
Experiment 4	.638 (.030)	.700 (.028)	.231 (.021)	.272 (.030)
Paintings Line				
Experiment 1	.605 (.025)	.696 (.025)	.098 (.013)	.159 (.021)
Experiment 2	.596 (.029)	.602 (.042)	.249 (.027)	.229 (.028)
Experiment 3	.634 (.057)	.607 (.049)	.141 (.041)	.139 (.047)
Experiment 4	.651 (.035)	.740 (.045)	.170 (.031)	.155 (.020)
Experiment 5: Motivated	.689 (.028)	.689 (.021)	.144 (.016)	.131 (.016)
Experiment 5: Non-Motivated	.657 (.025)	.678 (.026)	.113 (.017)	.146 (.019)
Poetry Line				
Experiment 1	.722 (.025)	.766 (.020)	.175 (.028)	.310 (.028)
Experiment 2	.635 (.037)	.781 (.031)	.176 (.029)	.158 (.040)
Experiment 3	.706 (.023)	.730 (.031)	.162 (.025)	
Experiment 4: Human Voices	.644 (.029)	.801 (.021)	.080 (.019)	
Experiment 5: Robot Voices	.612 (.029)	.735 (.024)	.213 (.016)	