Psychonomic Bulletin & Review
2009, 16 (1), 88-92
doi:10.3758/PBR.16.1.88

Surprising feedback improves later memory

Lisa K. Fazio and Elizabeth J. Marsh
Duke University, Durham, North Carolina

The hypercorrection effect is the finding that high-confidence errors are more likely to be corrected after feedback than are low-confidence errors (Butterfield & Metcalfe, 2001). In two experiments, we explored the idea that the hypercorrection effect results from increased attention to surprising feedback. In Experiment 1, participants were more likely to remember the appearance of the presented feedback when the feedback did not match expectations. In Experiment 2, we replicated this effect using more distinctive sources and also demonstrated the hypercorrection effect in this modified paradigm. Overall, participants better remembered both the surface features and the content of surprising feedback.

People do not have perfect knowledge about the world around them. As we go about our lives and interact with the world, we discover errors in our knowledge that we have to correct. How do we correct these errors? An examination of which false beliefs are easier versus more difficult to update should inform us about the mechanisms of correction. Many theories of memory hold that errors that one believes in very strongly—those made with high confidence—are the most difficult to correct later (see, e.g., McGeoch, 1942; Raaijmakers & Shiffrin, 1981). The argument is that errors made with high confidence are firmly established in our memories and are thus difficult to eradicate from our knowledge base.

Intriguingly, several studies have shown that high-confidence errors are actually more likely to be corrected after feedback than are low-confidence errors. In an early demonstration, participants read short paragraphs about the eye and then answered multiple-choice questions, rated their confidence in each answer, and received feedback about the correct answers (Kulhavy, Yekovich, & Dyer, 1976). On a final multiple-choice test that repeated the same 30 questions, participants corrected more of their high-confidence errors than their low-confidence errors. More recently, Butterfield and Metcalfe (2001) found the same effect with different stimuli. In their experiment, participants answered general knowledge questions such as "What poison did Socrates take at his execution?" Participants rated their confidence in each response and then were told the correct answer to each question. Similar to Kulhavy et al., Butterfield and Metcalfe (2001) found that high-confidence errors were more likely to be corrected on a retest than were low-confidence errors. The authors named this finding the *hypercorrection effect*.

Why is it that these high-confidence errors, which should be firmly established in memory and difficult to update, are instead more likely to be corrected than are low-confidence errors? One possibility is that participants attend more to unexpected feedback, with positive consequences for memory. In other words, when a participant makes an error with high confidence, the feedback is surprising, leading the learner to encode the feedback more deeply. This hypothesis is similar to Kulhavy's (1977; Kulhavy et al., 1976) model of how feedback affects learning, and it owes a debt to Rescorla and Wagner's (1972) model of animal learning (which stated that learning occurs fastest when events violate the organism's expectations). Kulhavy proposed that a large discrepancy between the participant's initial beliefs and the correct answer leads the participant to expend more effort to correct the misunderstanding.

One prediction of this model is that participants will choose to spend more time studying the feedback after a high-confidence error; this was confirmed by Kulhavy (1977). The hypercorrection effect occurs even when the duration of the feedback is held constant (as in Butterfield & Metcalfe, 2001), however, so the challenge is to find evidence for surprise when differential study times are not possible. Some support comes from neuroimaging data; for example, Butterfield and Mangels (2003) used ERPs to show that high-confidence errors elicited activity in frontal areas that have been linked to novelty in other studies (see Butterfield, 2005, for a similar result using fMRI).

When one answers questions, the feedback might be surprising in two different situations. In addition to being surprised by high-confidence errors, an individual should also be surprised when he or she responds with a guess (and rates confidence as low) and finds out that it was correct. A test–feedback–retest paradigm allows only for examination of errors, since low-confidence correct answers do not need to be corrected, but the surprise hypothesis predicts that both situations should have consequences for

L. K. Fazio, lkf@duke.edu

attention and, in turn, for later memory. Butterfield and Metcalfe (2006) used this logic in a pair of experiments in which participants did a tone-detection task while answering the initial general knowledge questions and then were retested with full attention. During the initial test, participants simply had to press a key whenever they heard a tone; critical was participants' ability to detect tones that were played concurrently with feedback. Surprising feedback was presumed to divert attention from the tone-detection task. Consistent with this assumption, participants missed more tones when the feedback revealed an error that had been made with high confidence. With correct answers, tone detection was better for high-confidence responses than for correct guesses. Overall, tone detection was negatively related to performance on the retest, suggesting that participants encoded the feedback at the expense of detecting the tone.

Our research also takes advantage of the fact that the surprise hypothesis predicts increased attention (and memory) for both high-confidence errors and low-confidence correct answers. Instead of using distraction from another task to infer that there was attention to the feedback, however, we chose a more direct measure of attention to the feedback, one that could be measured for both correct answers and errors: memory for the feedback's appearance. This dependent measure has been used in emotion research, with the result that memory is better for the surface features (e.g., font colors) of attention-grabbing emotional and taboo words than for the surface features of neutral words (Doerksen & Shimamura, 2001; MacKay & Ahmetzanov, 2005). This is a measure of source memory, or memory for the "conditions under which a memory is acquired" (Johnson, Hashtroudi, & Lindsay, 1993, p. 3). We are using a broad definition of source memory that includes everything that gets encoded about the feedback other than its content. Our argument is that source memory is better when feedback is surprising, for both correct and incorrect answers.

In Experiment 1, participants answered general knowledge questions, rated their confidence in each answer, and received feedback in the form of the correct answer to each question. Critically, the feedback appeared in either red or green font. After a short delay, participants identified whether each correct answer had been presented in green or red during the feedback phase. If a discrepancy between the participant's expectation and the feedback led to a deeper encoding of the feedback, then source memory should have been better for high-confidence errors *and* low-confidence correct responses than for low-confidence errors or confident correct answers. In other words, when the feedback confirmed participants' beliefs, they should have paid less attention to it, resulting in lower memory for the feedback's appearance. This same relationship was expected in Experiment 2, in which either a male or a female voice delivered the feedback. One group of participants completed a source test (as in Experiment 1): For each correct answer, they identified whether it had been spoken in a male or a female voice. Other participants were simply retested on the general knowledge questions. Experiment 2 was thus designed to generalize the relationship between confidence and appearance memory to a different source judgment, as well as to demonstrate the standard hypercorrection effect in our modified paradigm. Because of the similarities between the two experiments, they will be discussed together in a single general discussion.

EXPERIMENT 1

Method

Participants. Forty-six Duke University undergraduates participated in the experiment for partial fulfillment of a course requirement. Seventeen additional participants were tested, but they performed at chance on the source discrimination task; thus, their data were excluded from the analyses. "Chance" was defined as answering less than 55% of the source questions correctly. None of the participants was color blind.

Materials. One hundred forty general knowledge questions were selected from the Nelson and Narens (1980) norms. Questions ranged in difficulty; on average, 40% of participants in the norming study answered these items correctly (correct response rates ranged across items from 0% to 92%). The feedback appeared in Times New Roman font. It was either red and italicized in 64-point font, or green, underlined, and bolded in 12-point font. All other text was presented in light-blue 24-point font.

Procedure. The experiment began with a general knowledge test. Participants were told to answer a series of questions and rate their confidence in each answer. They were warned that some of the questions would be difficult and that they should make educated guesses, or else respond, "I don't know." Furthermore, they were told they would receive feedback on their answers and that they would later take a second test. Critically, the nature of the second test was never mentioned.

Participants typed their response to each question, and they rated their confidence using a 7-point scale. Following Butterfield and Metcalfe (2001), the scale ranged from 1 (*sure wrong*) to 4 (*unsure*) to 7 (*sure correct*). The correct answer appeared for 5 sec after each confidence rating was recorded. This feedback took the form of a sentence, and it was presented regardless of whether or not the question was answered correctly. For example, if the question was "What's the longest river in South America?" then the feedback was "Amazon is the longest river in South America." For half of the items, the feedback was presented in the red font, whereas for the other half the feedback appeared in the green font.

Immediately following the general knowledge test, participants completed a source test on their memory for the feedback's appearance. The feedback sentences were tested one at a time, in random order, in the light-blue font. For each item, participants identified whether the feedback had been presented previously in red or green font. After the source test, participants were debriefed and were thanked for their participation.

Results

Unless otherwise noted, differences were significant at the .05 level.

Initial test. On the initial test, participants answered an average of 43% of the questions correctly, and their average confidence was 4.11. Participants were well calibrated in their use of the confidence scale; the average within-subjects gamma correlation between initial test accuracy and confidence was .78.

Source test. Participants correctly identified the prior color of the feedback for 69% of the facts.

Our primary interest was in the relationship between confidence on the initial test and performance on the source test. The surprise hypothesis predicts a different relationship between confidence and source memory for

items that are answered correctly versus those that are answered incorrectly on the initial general knowledge test. For general knowledge questions that are answered *correctly*, the feedback would be unexpected for guesses, thus leading to better source memory for low-confidence correct answers than for high-confidence ones. In contrast, for general knowledge questions that are answered *incorrectly*, the feedback would be surprising for high-confidence errors, thus leading to better source memory for high-confidence errors than for low-confidence ones. In short, the surprise hypothesis predicts a negative relationship between source memory and confidence for items that are answered correctly on the initial general knowledge test, but a positive relationship for items that are answered incorrectly.

Figure 1 shows the relationship between source memory and confidence as a function of correctness on the initial general knowledge test. As predicted, source memory was highest when participants' confidence was mismatched with the accuracy of their original responses. For correct answers, lower confidence on the initial test was associated with better source memory. The mean within-subjects gamma correlation between initial confidence and later source memory was significantly negative [$\gamma = -.19$, $t(45) = 2.61$, $SEM = .07$]. For incorrect answers, higher confidence was associated with better source memory [$\gamma = .12$, $t(45) = 2.23$, $SEM = .05$].

A series of additional analyses were conducted to ensure that the key results were not due to differential memory for the red font ($M = .72$), which turned out to be more memorable than the green font ($M = .65$) [$t(45) = 3.07$, $SEM = .02$]. Half of the feedback statements were presented in red and half in green, but because we could not predict a priori an individual's responses nor their confidence in these responses, we could not counterbalance the font across the 14 cells. For errors, critically, red and green feedback were not unequally distributed across the seven levels of confidence [$F(6,270) = 1.05$, $MS_e = 2.65$, $p > .3$]; thus, better memory for red feedback could not explain the positive relationship between confidence and source memory that was found for errors. For correct responses, disproportionately more red feedback occurred in the high-confidence cells [$F(6,270) = 6.95$, $MS_e = 2.42$], but this is not concerning, since the result that was predicted (and obtained) was in the opposite direction.

In short, the results were consistent with the surprise hypothesis: The relationship between confidence and source memory was positive for errors but negative for correct answers.

Because Experiment 1 focused on source memory, there was no measure of error correction. That is, a second general knowledge test was not administered after the source test, since the source memory test effectively presented the feedback for a second time. In Experiment 2, one group of participants took the source test and another group was retested on the general knowledge questions in order to ensure that the changed paradigm did not eliminate the basic hypercorrection effect. In addition, the sources were made more distinctive in Experiment 2 to minimize the loss of participants because of chance performance on the source test.

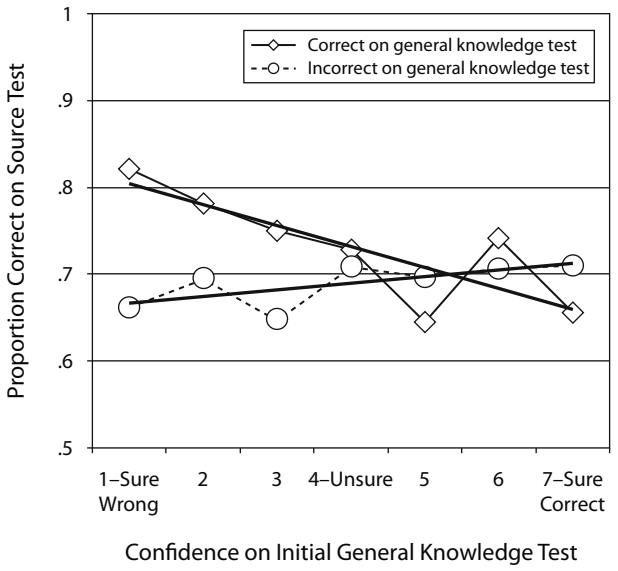

Figure 1. Average proportion correct on the source test for each confidence level in Experiment 1, as a function of whether the answer on the initial test was correct or incorrect. The bolded lines are the best-fitting trend lines.

EXPERIMENT 2

Method

Participants. Seventy-two undergraduates participated in the experiment for partial fulfillment of a course requirement. In the final test phase, 50 participants took the source test. Six additional participants were tested in this condition but were excluded because they performed at chance on the source test (chance was defined in the same terms as it was in Experiment 1). Twenty-two participants were in the retest condition; 1 additional participant was tested but was excluded because he corrected all of his initial errors on the second test, making it impossible to calculate the relationship between his confidence in the initial error and the probability of it being corrected on the retest.

Materials. We used 120 of the original 140 questions from Experiment 1. Feedback was presented in one of two ways. For half of the items, a female voice read the feedback aloud while a woman's picture appeared on the left side of the computer screen and the feedback, printed in pink lettering, appeared on the right. For the other half of the items, the voice was male and the computer screen showed a man's picture on the right and the feedback, in blue lettering, on the left.

Procedure. As in Experiment 1, participants answered a series of general knowledge questions, rated their confidence in each response, and then received feedback. To improve source memory, the feedback appeared for 6 sec instead of 5 sec (as was the case in Experiment 1).

After the general knowledge test, participants in the source memory condition immediately began the source test. The feedback sentences appeared on the screen in a neutral font, and the participants identified whether the male or the female source had presented the feedback.

Participants in the retest condition solved visuospatial puzzles for 4 min before taking their final test, since pilot testing showed that participants were at ceiling without a short filler task. These partici-

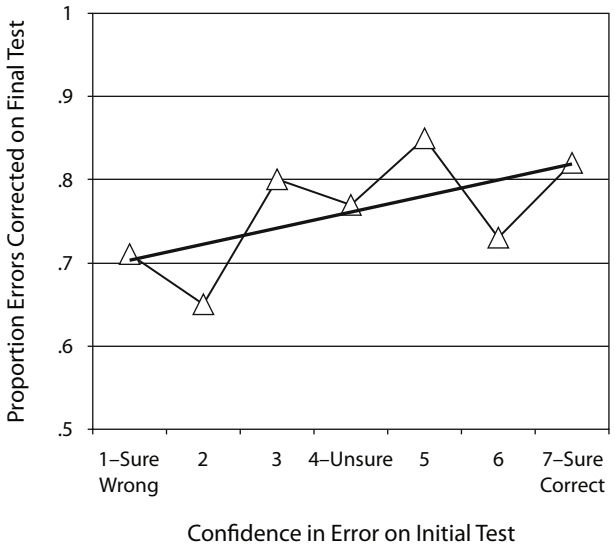

Figure 2. Average proportion of the errors on the first test that were corrected on the second test for each confidence level in Experiment 2. The bolded line is the best-fitting trend line.

pants then retook the general knowledge test, which was identical to the first test except that no feedback was provided.

Results

Unless otherwise noted, differences were significant at the .05 level.

Initial test. Performance on the initial test did not differ across the two conditions. Participants correctly answered 42% of the initial questions in the source condition and 43% in the retest condition ($t < 1$). Confidence on the initial test was also similar across the conditions, averaging 4.01 in the source condition and 4.21 in the general knowledge retest condition ($t < 1$). These values are similar to those observed in Experiment 1, as were participants' confidence–accuracy correlations. The average within-subjects gamma correlation between proportion correct and confidence on the initial test was .81 in the source condition and .76 in the retest condition [$t(70) = 1.35$, $SEM = .04$, $p = .18$].

General knowledge retest. For the participants in the general knowledge retest condition, we compared performance on the initial test with performance on the final test. Feedback improved performance, with participants answering 80% of the questions correctly on the second test, as compared with 43% on the initial test [$t(21) = 26.17$, $SEM = .01$].

Of primary interest was whether the hypercorrection effect occurred. For each of the seven confidence levels on the first test, we examined the proportion of errors that were successfully corrected on the second test. Figure 2 shows hypercorrection: Participants corrected more of the errors that had been committed with high confidence than those that had been made with low confidence. The mean within-subjects gamma correlation between initial confidence and proportion of errors later corrected was significantly positive [$\gamma = .23$, $t(21) = 2.27$, $SEM = .10$].

Source test. For the participants in the source condition, we examined memory for the source of the feedback. On average, participants correctly identified the source for 68% of the facts.

As in Experiment 1, our primary interest was in the relationship between confidence on the initial test and later memory for the source of the feedback. As we found in Experiment 1, there was a negative relationship between confidence and source memory for correct answers and a positive relationship between confidence and source memory for errors, as shown in Figure 3. After answering a question correctly, participants were more likely to remember the source of the feedback if they had answered with low confidence than if they had answered with high confidence [$\gamma = -.28$, $t(49) = 4.24$, $SEM = .07$]. The pattern was opposite for errors; participants were more likely to remember the source of the feedback if they had answered with high confidence than if they had answered with low confidence [$\gamma = .12$, $t(49) = 2.18$, $SEM = .06$].

As in Experiment 1, we conducted additional analyses to ensure that our results were not due to one source being more memorable than the other. We found that the male source was more likely to be accurately identified ($M = .70$) than the female source ($M = .66$) [$t(49) = 2.53$, $SEM = .02$], but the male and female feedback was not unequally distributed across the confidence levels for correct answers [$F(6,294) = 1.80$, $MS_e = 2.29$, $p > .1$] or for errors ($F < 1$); thus, better memory for the male source cannot explain our results.

DISCUSSION

In two experiments, surprising feedback improved memory for both the surface features and the content of presented feedback. In Experiment 1, participants were

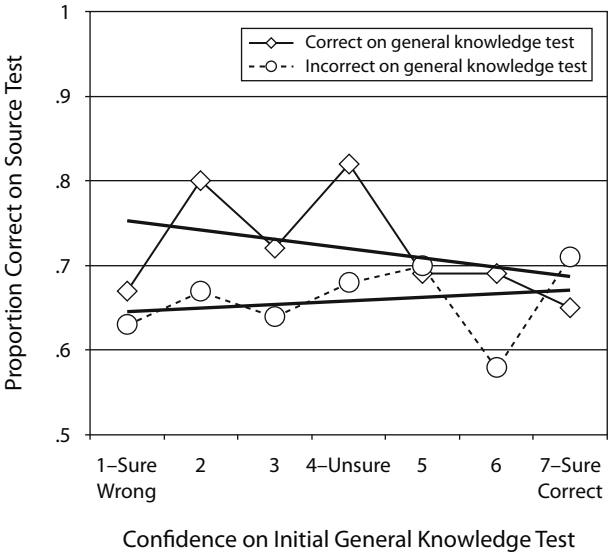

Figure 3. Average proportion correct on the source test for each confidence level in Experiment 2, split by whether the answer on the initial test was correct or incorrect. The bolded lines are the best-fitting trend lines.

better able to remember the color of feedback when it was incongruent with their expectations. That is, source memory was better for feedback that had been presented in response to both correct guesses and errors made with high confidence. In Experiment 2, participants showed improved memory for both the content and the source of the feedback. Participants were more likely to correct high-confidence errors than low-confidence errors, and they were more likely to remember the source of the feedback when it was unexpected.

Although the observed relationships between initial confidence and source memory were relatively small, they occurred as predicted in both experiments and for both correct and erroneous answers. It is not surprising that the effects were small, given that remembering the appearance of the feedback was not the participants' main task. The participants in both experiments were led to believe that they would be retested on the general knowledge questions—the source memory test was unexpected; thus, most of the participants' additional attention should have been, and was, directed toward the content of the surprising feedback, rather than toward its surface features. This can be seen most clearly in Experiment 2, in which memory for the content of the feedback (the correct answer) increased more than 10% across the confidence levels, whereas source memory increased less than 5%.

These experiments support the surprise hypothesis, which states that unexpected feedback leads to a greater expenditure of effort to encode that feedback, with positive consequences for memory. Data across laboratories are converging in support of the surprise hypothesis. When these results are put together, a consistent picture emerges: Feedback can be surprising (Butterfield & Mangels, 2003), leading to a focus on the feedback (the present study) at the expense of other tasks (Butterfield & Metcalfe, 2006).

In addition to the surprise hypothesis, there is at least one other possible explanation of the hypercorrection effect. The *knowledge hypothesis* posits that confidence tends to be correlated with how much a participant knows generally about the target domain (Butterfield & Metcalfe, 2001). The argument is that if participants have little knowledge about a domain, they have nothing with which to associate the incoming information. In other words, it is more difficult to integrate the correct answer into their semantic memory if it is in an unfamiliar domain. Although the present experiments were not designed to test the knowledge hypothesis, it is not immediately clear what the knowledge hypothesis would predict about memory for the source of the feedback. In particular, we doubt that the knowledge hypothesis would predict a negative relationship between source memory and confidence in correct answers. Of course, our data do not rule out the knowledge hypothesis, since the two hypotheses are not mutually exclusive. It is quite plausible that knowledge updating requires both deep encoding of the feedback and a knowledge structure that allows the new information to be easily assimilated, but our data suggest that differences in domain knowledge are unlikely to be solely responsible for the hypercorrection effect.

It is important for people to be able to accurately update their general knowledge. We believe that confidence judgments play an important role in dictating which errors are most essential to correct. Because confidence judgments are in general a valid indicator of overall accuracy (Brewer, Sampaio, & Barlow, 2005; Perfect, Watson, & Wagstaff, 1993), it is informative when there is a conflict between a person's confidence and the actual answer. The contradictory information tells the person that something is seriously wrong with his or her knowledge structure. It is the importance of this miscalibration that causes the feedback to be better processed and better remembered.

AUTHOR NOTE

This work was supported by a collaborative activity award from the James S. McDonnell Foundation. We thank Barbie Huelser for help with data collection. Correspondence concerning this article should be addressed to L. K. Fazio, Department of Psychology and Neuroscience, Duke University, 9 Flowers Drive, Box 90086, Durham, NC 27708-0086 (e-mail: lkf@duke.edu).

REFERENCES

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory & Language*, **52**, 618-627.

Butterfield, B. A. (2005). The hypercorrection effect and its neural correlates. *Dissertation Abstracts International: Section B: The Sciences & Engineering*, **66**, 2846.

Butterfield, B. [A.], & Mangels, J. A. (2003). Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research*, **17**, 793-817.

Butterfield, B. [A.], & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1491-1494.

Butterfield, B. [A.], & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition & Learning*, **1**, 69-84.

Doerksen, S., & Shimamura, A. P. (2001). Source memory enhancement for emotional words. *Emotion*, **1**, 5-11.

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

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, **47**, 211-232.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology*, **68**, 522-528.

MacKay, D. G., & Ahmetzanov, M. V. (2005). Emotion, memory, and attention in the taboo Stroop paradigm: An experimental analogue of flashbulb memories. *Psychological Science*, **16**, 25-32.

McGeoch, J. A. (1942). *The psychology of human learning, an introduction*. New York: Longmans, Green.

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

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology*, **78**, 144-147.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, **88**, 93-134.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

(Manuscript received May 15, 2008;
revision accepted for publication August 10, 2008.)