DOCUMENT RESUME

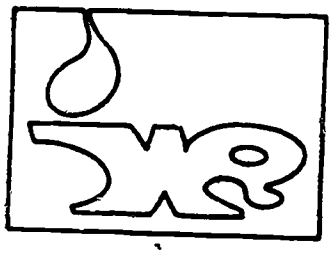
ED 273 935 CS 008 548

AUTHOR Glenberg, Arthur M.; And Others
TITLE Enhancing Calibration. Program Report 86-14.
INSTITUTION Wisconsin Center for Education Research, Madison.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE 1 Oct 86
CONTRACT N0014-85-K-0063
NOTE 53p.
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluation Criteria; Higher Education; *Learning Processes; Learning Strategies; *Reading Comprehension; *Reading Research; *Reading Tests; *Testing Problems; Test Reliability
IDENTIFIERS *Calibration

ABSTRACT

Defining calibration of comprehension as the correlation between subjective assessments of knowledge gained from reading and performance on an objective test, this paper draws from literature, as well as the findings of original experiments, to examine issues related to the subject . The paper first documents the claim that poor calibration is the rule rather than the exception, and that poor calibration is also typical in at least one other domain, problem solving. It suggests that the high levels of calibration reported in studies on the calibration of probabilities and feeling-of-knowing research may be dependent on using feedback from taking the test to assess the probability of correct performance on the test. The paper next presents two experiments demonstrating that poor calibration is not associated with a particular type of performance test, but is found with inference tests, verbatim recognition tests, and idea recognition tests. It further shows that poor calibration cannot be attributed to unreliable testing procedures. The paper then offers evidence from three experiments indicating that a likely reason for poor calibration is that subjects assess familiarity with the general domain of a text instead of assessing knowledge gained from a particular text. Next, it demonstrates that calibration of comprehension can be enhanced if subjects are given a pretest that provides self-generated feedback--but that even this ability is limited. The paper concludes with a discussion of the implications of these findings for theories of representation of knowledge gaining from reading. Sample materials used in the experiments are appended. (Author/FL)



Program Report 86-14
October 1, 1986

Enhancing Calibration

by Arthur M. Glenberg,
Thomas Sanocki, William Epstein,
and Craig Morris

This work was supported by Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N0014-85-K-0063, Contract Authority Identification No. NR 702-012. Approved for public release; distribution unlimited. Reproduction in whole or part is permitted for any purpose of the United States Government.

Wisconsin Center for Education Research
School of Education, University of Wisconsin-Madison

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approval for public release; distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) WCER Program Report 86-14	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Wisconsin Center for Education Research 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Personnel Training Research Programs Office of Naval Research (Code 1142PT)
6c. ADDRESS (City, State, and ZIP Code) 1025 West Johnson Street Madison, WI 53706	7b. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N0014-85-K-0644
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61153N	RR04206	RR04206-0C	NR702-012

11. TITLE (Include Security Classification)
Enhancing Calibration (Unclassified)

12. PERSONAL AUTHOR(S)
Glenberg, A. M., Sanocki, T., Epstein, W., and Morris, C.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final Report	FROM 9/85 TO 9/86	86-10-1	50

16. SUPPLEMENTARY NOTATION
Under review

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
05	10		comprehension, meta-comprehension

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Calibration of comprehension is the correlation between subjective assessments of knowledge gained from reading and performance on an objective test. Contrary to intuition, typically this correlation is close to zero. This article is structured around four points concerning calibration of comprehension. First, poor calibration is the rule, rather than the exception. It has been repeatedly demonstrated in our laboratory and in others. Poor calibration is also typical in at least one other domain, problem solving. The high levels of calibration reported in studies on the calibration of probabilities and feeling of knowing research may be dependent on using feedback from taking the test to assess the probability of correct performance on the test.

Second, we present two experiments that demonstrate that poor calibration is not associated with a particular type of performance test, but it is found with inference tests, verbatim recognition tests, and idea recognition tests. For the most part, poor calibration is found when the test is given immediately after reading as well as when the test is given after a delay. Also, we demonstrate that poor calibration cannot be attributed to

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Dr. Michael Shafto	(202) 696-4596	ONR 1142PT

19, ABSTRACT (continued)

unreliable testing procedures.

Third, the evidence from three experiments indicates that a likely reason for poor calibration is that subjects assess familiarity with the general domain of a text instead of assessing knowledge gained from a particular text. Assessing domain familiarity is probably easier than assessing knowledge gained from a particular text. Also, under some conditions, applying a domain familiarity strategy does result in spurious calibration, thereby reinforcing application of the strategy.

Fourth, we demonstrate that calibration of comprehension can be enhanced if subjects are given a pre-test that provides (self-generated) feedback. Even this ability is limited, however. Calibration is only enhanced when the processes and knowledge tapped by the pre-test are closely related to the processes and knowledge required on the criterion test. Under these conditions, subjects apparently use feedback from the pre-test to predict criterion test performance with a modest degree of accuracy. We briefly discuss the implications of these results for theories of representation of knowledge gained from reading.

Program Report 86-14

Enhancing Calibration

Arthur M. Glenberg, Thomas Sanocki, William Epstein,

and Craig Morris

Wisconsin Center for Education Research

School of Education

University of Wisconsin-Madison

October 1, 1986

This work was supported by Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N0014-85-K-0063, Contract Authority Identification No. NR 702-012. Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Abstract

Calibration of comprehension is the correlation between subjective assessments of knowledge gained from reading and performance on an objective test. Contrary to intuition, typically this correlation is close to zero. This article is structured around four points concerning calibration of comprehension. First, poor calibration is the rule, rather than the exception. It has been repeatedly demonstrated in our laboratory and in others. Poor calibration is also typical in at least one other domain, problem solving. The high levels of calibration reported in studies on the calibration of probabilities and feeling of knowing research may be dependent on using feedback from taking the test to assess the probability of correct performance on the test.

Second, we present two experiments that demonstrate that poor calibration is not associated with a particular type of performance test, but it is found with inference tests, verbatim recognition tests, and idea recognition tests. For the most part, poor calibration is found when the test is given immediately after reading as well as when the test is given after a delay. Also, we demonstrate that poor calibration cannot be attributed to unreliable testing procedures.

Third, the evidence from three experiments indicates that a likely reason for poor calibration is that subjects assess familiarity with the general domain of a text instead of assessing knowledge gained from a particular text. Assessing domain familiarity is probably easier than assessing knowledge gained from a particular text. Also, under some conditions, applying a domain familiarity strategy does result in spurious calibration, thereby reinforcing application of the strategy.

Fourth, we demonstrate that calibration of comprehension can be enhanced if subjects are given a pre-test that provides (self-generated) feedback. Even this ability is limited, however. Calibration is only enhanced when the processes and knowledge tapped by the pre-test are closely related to the processes and knowledge required on the criterion test. Under these conditions, subjects apparently use feedback from the pre-test to predict criterion test performance with a modest degree of accuracy. We briefly discuss the implications of these results for theories of representation of knowledge gained from reading.

In preparing for a test of learning, a rational strategy is to study until one believes that the material is learned. Studying for less time is risky; studying for more time may be wasteful. For this strategy to be effective, however, beliefs and judgements about how much has been learned must be calibrated. That is, these beliefs must be correlated with performance. Unfortunately for learners, calibration of comprehension often is close to zero.

We have four goals for this article. The first is to document our claim that beliefs about how much has been learned are often uncorrelated with performance on a test of comprehension. Second, we will demonstrate that the lack of correlation is not due to some methodological artifact, but that it is representative of a wide range of situations. Third, we present data supporting one general account for the lack of correlation. Finally, we will demonstrate a method for enhancing calibration.

In total, we believe that these results have implications for understanding meta-cognitive processes and comprehension of expository text. These implications depart in at least two significant ways from standard theorizing in the field. To preview, our subjects seem to form representations that are specific rather than abstract; also, these representations do not seem to be well-organized.

Readers are Poorly Calibrated

We define calibration of comprehension as the correlation between ratings of confidence in comprehension and actual performance on an objective test of comprehension. A correlation near 1.0 indicates very good calibration; a correlation near zero indicates little calibration. The general finding across a variety of procedures is that calibration is near zero.

Data from our own laboratory has almost uniformly demonstrated poor calibration. Glenberg, Wilkinson, and Epstein (1982) and Epstein, Glenberg, and Bradley (1984) used a contradiction procedure. Subjects read expositions with the explicit instruction to find sentences embedded in the text that were contradictory. Subjects frequently reported high confidence in their understanding of the text after failing to find contradictions between adjacent sentences. This mismatch between confidence and performance is indicative of poor calibration.

Glenberg and Epstein (1985) measured calibration more directly. After reading a number of (unadulterated) brief expositions, subjects rated confidence in ability to verify inferences derived from the texts. While making the confidence judgement for a text, the subject had available the specific principle that would be used to draw the inference. Nonetheless, the correlation between confidence and performance was not significantly different from zero. Furthermore, the correlation did not improve with practice, nor did it improve when the confidence judgement was elicited immediately before the inference verification test for each passage.

In a more recent report, Glenberg and Epstein (in press) examined calibration as a function of expertise in a domain of knowledge. Students with

a wide range of experience in physics or music read texts in music theory and physics. After reading, the students provided confidence assessments for each text and answered inference questions for each text. As expected, music students were more confident on the music texts than the physics texts, and their performance on the music inference questions was better than their performance on the physics inference questions. Analogous results were found for the physics students. Thus, across domains of knowledge, these students were calibrated. Nonetheless, within a domain, there was essentially no calibration. Furthermore, expertise in the domain either was uncorrelated with calibration (for the music students), or was negatively correlated with calibration (for the physics students).

Maki (Maki & Berry, 1984; Maki & Monson 1985) has used a procedure similar to the calibration procedure. Although her results are somewhat complicated, the overall picture is of very poor calibration. Subjects in Maki and Berry (1984) read a chapter from a psychology textbook, rated confidence in their future test performance, and then took a test one day later. Subjects who performed above the median on the test had a modest amount of calibration (r =.15). Those who performed below the median were not calibrated (r =-.03). In a second experiment, subjects were given an immediate test over the first and second parts of the chapter. On the first half of the chapter the average correlation for all subjects was .23. On the second half of the chapter, however, the average correlation was essentially zero.

In Maki and Monson (1985), subjects read two chapter halves and the second half was read either once, twice in a massed fashion, or twice in a distributed fashion. Although distribution of study affected test performance, it did not affect calibration. Furthermore, calibration was very poor. For subjects above the median on test performance, r =.12, for subjects below the median r =.06.

Moving away from the literature on calibration for text, Metcalfe (1986) contrasted calibration for memory with calibration for problem solving. For memory calibration, subjects predicted how well they would recognize answers to trivia-like questions that could not be recalled. The (gamma) correlations ranged from .45 to .52. These same subjects were also given "insight" problems to solve. For problems the subjects could not solve immediately, the subject provided a rating as to the likelihood of success given an additional five minutes to work on the problem. The correlations between the ratings and problem solving performance ranged from -.32 to .10. Thus subjects were poorly calibrated in the problem solving domain.

Even within the memory domain, the relatively high correlations between confidence and performance may not require a very impressive ability to assess knowledge. The usual interpretation is that the correlations reflect some form of privileged access (e.g., Lovelace, 1984) to knowledge, as implied by the term "feeling of knowing." Alternatively, the correlations might reflect the use of public knowledge, for example, that certain types of problems are difficult. Nelson et al. (1986) demonstrated that an individual's predictions were not as highly correlated with performance as was normative item difficulty. There was some evidence for privileged access, but not to an impressive degree.

Vesonder and Voss (1985) also demonstrated that feeling of knowing judgements may be based more on public knowledge than on accurate assessments of

private knowledge. In their second experiment, Learners studied sentences for recall and predicted performance. Observers viewed the same sequence and predicted the Learner's performance. Overall, the Learner's predictions were somewhat more accurate than the Observer's predictions. Nevertheless, on the subset of stimuli missed after the first study trial, the predictions made by the Learner were not more accurate than those made by the Observer.

In summary, the evidence indicates that the ability to self-assess comprehension (Glenberg and Epstein, 1985; Maki and Berry, 1984) and problem-solving (Metcalfe, 1986) is not impressive. Additionally, although feeling of knowing predictions can be accurate, it is not clear that these predictions are based on priviliged access to an individual's specific knowledge.

Calibration of Comprehension is Poor for Three Different Types of Tests, and at Two Different Retention Intervals

The evidence adumbrated seems to be contradicted by our intuitions. When we read, it seems plain enough when we understand and when we don't. If these intuitions are correct, then the evidence implies not a problem with self-assessment of comprehension, but a problem with the procedures used to measure the accuracy of self-assessment. In this section we examine two possible problems with the procedures used in Glenberg and Epstein (1985, in press).

The standard procedure has been to use an inference verification test. The subject assesses ability to judge whether inferences are correctly drawn from a principle. The principle has been encountered in a previously read text, and it is available while the assessment is being made. Our reasoning behind use of this task is threefold. First, ability to draw inferences seems to be a more reasonable measure of understanding than recall or recognition. Second, if a subject has knowledge about a principle, then inference verification should be more accurate than if the subject has no knowledge regarding the principle. Finally, if a subject has access to that knowledge, predictions should reflect performance.

Nonetheless, accurate assessment of performance on the inference verification task may be quite difficult. First, a variety of inferences can be drawn using the principle, and the subject may not be able to properly assess ability in such a wide domain. Second, inference verification must require types of knowledge (e.g., logical rules) quite distinct from the principle. Thus assessments of knowledge of the principle (e.g., recallability) may be accurate, but not predict performance on the inference verification test which requires application of the principle in a reasoning task. To examine this possibility, we designed experiments using a variety of tests, including inference verification, verbatim recognition, and recognition of ideas from the text.

A second possible problem with the standard procedure concerns the time of testing relative to the time of reading. Glenberg and Epstein (1985) demonstrated that placement of the test relative to the confidence assessment did not affect calibration. Maki and Berry (1984) did demonstrate changes in calibration with a delay, but their manipulation confounded the time between

reading and the test with the time between the confidence assessment and the test.

Confidence assessments might make use of information that is valid only within a limited temporal range. For example, subjects may formulate an accurate assessment of comprehension while reading, and simply recall that assessment (rather than performing a re-assessment) to make the confidence judgment. Although the assessment may be accurate shortly after reading, it may become less valid over a longer retention interval, because the subject is likely to forget some of the read material.

Experiment 1: Inference Verification and Verbatim Recognition Tested Immediately and After a Delay

Two variables were manipulated in this factorial experiment. The first was type of test: Half of the subjects received inference verification tests and half received verbatim recognition tests. The verbatim recognition test consisted of a pair of sentences that were close paraphrases. The subject's task was to choose the sentence that was a verbatim reproduction from the text. Examples of these materials are included in Appendix A.

The second variable was the delay between reading the passage and the comprehension test. In the immediate condition, each passage was followed immediately by the confidence assessment for that passage and then the test. In the delayed condition, the subject read through all 15 passages. Then the subject was presented with 15 pairs consisting of a confidence assessment and associated test.

Both variables were manipulated between subjects. Also, subjects were fully informed as to the type of test they would receive and the delay between reading and testing.

Method

Subjects. The subjects were 80 students attending the summer session at the University of Wisconsin-Madison who were paid for participating. Twenty subjects were randomly assigned to each of the four groups formed by the factorial combination of the two independent variables. Subjects were run in groups of 1-8 individuals.

Materials. We wrote 15 texts (and three practice texts) on various topics. Versions of these texts were used in Glenberg and Epstein (1985). Each text was (a) one paragraph long, and (b) written to illustrate, exemplify, or amplify a central principle that was stated explicity in the text. A paraphrase of the central principle was also prepared. Half the texts contained the original statement of the principle and half the paraphrase. Additionally, associated with each text were two inference verification tests. One test was an inference derivable from the central principle (true inference), the other was a contradiction of a true inference that could not be derived from the central principle (false inference, see Appendix A).

The confidence assessment form was headed with the title of a specific text. For subjects receiving the inference verification test, the confidence

assessment indicated that the subject should "use the following scale to report your confidence that you are able to use what you have learned in this text to draw correct inferences" regarding the central principle of the text. That principle then appeared on the form above a six-point confidence scale. The number one on the scale was labeled "very low", and the number six was labeled "very high". For subjects receiving the verbatim recognition test, the confidence assessment form indicated that the subject should use the "scale to report how confident you are that you will be able to choose a verbatim (word for word) sentence from the text when given a choice between a verbatim sentence and a paraphrase (restatement of the sentence)." In either case, the corresponding test appeared on the next page of the subject's booklet.

Three practice texts were constructed along the same lines as the experimental texts. The confidence assessments and tests associated with the practice passages reflected the conditions the subject would experience during the main part of the experiment (inference verification or verbatim recognition, and immediate or delayed testing).

All materials were collated into individual booklets for each subject. A separate page was used for each text, confidence assessment, and test. In the immediate condition, each text was followed by the corresponding confidence assessment and test on the next two pages. In the delayed condition, the 15 texts were on consecutive pages (with the order randomized for each subject), and the 15 pairs of confidence assessments and tests followed the last text (in the same order as the texts). A subject was free to proceed through the booklet at his or her own pace. The only constraint was that once a page was turned, it could not be turned back.

Results and Discussion

Insert Table 1

The results are in Table 1. Each dependent measure was analyzed using a two factor factorial analyses of variance with the probability of a type 1 error set at .05.

Confidence. Subjects were somewhat more confident for the inference verification test than for the verbatim recognition test, $F(1, 76) = 10.34$, $MSE = .48$. Also, there was a significant interaction so that the difference between the immediate and the delayed condition was larger for the verbatim recognition test than for the inference verification test, $F(1, 76) = 4.10$. The most important feature of the confidence data is that there is variability, thus a correlation between confidence and performance (calibration) is not artificially constrained by floor or ceiling effects.

Proportion correct. In general, subjects were more often correct on the immediate test than on the delayed test, $F(1, 76) = 4.71$, $MSE = .02$. The interaction between delay and type of test was significant, $F(1, 76) = 9.83$, however, indicating that the advantage for the immediate test was only for verbatim recognition. Once again, these data are not constrained by floor or ceiling effects.

Calibration. Calibration in comprehension is the correlation between confidence ratings and performance on the tests. A separate calibration coefficient can be computed for each subject by measuring the association between the 15 confidence scores (1-6) and the 15 performance scores (0 or 1). Subjects for whom there was no variability in either the confidence scores or the performance scores were eliminated from the analysis. Two correlation coefficients were computed for each subject. The first, rpb, is the point-biserial correlation and it can be interpreted as a Pearson product-moment coefficient. The second is the non-parametric gamma (G) that has been recommended for data of this sort (Nelson, 1984). It also ranges from -1 to 1 with zero indicating no relationship. One interpretation of G depends on considering pairs of texts that differ in both confidence and performance. Considering all of these pairwise comparisons, G is the difference between the probability that the text with the higher confidence is the correct one and the probability that the text with the lower confidence is the correct one.

There is a slight hint that calibration for inference verification was greater than calibration for verbatim recognition. Statistically, however, the effect was not significant for either measure of calibration. In fact, neither the main effects nor the interactions are statistically significant, nor are any of the calibration coefficients taken alone significantly different from zero.

One might object that the experiment lacks power, but we had sufficient power to detect some rather small effects in confidence and proportion correct. One might also object that the components of the calibration, based on but a single measure of confidence and a single measure of knowledge for each text, are unreliable, thus reducing calibration. We will demonstrate in the next experiment, however, that reliability is not a significant problem.

The most straightforward conclusion is that subjects were not calibrated. Furthermore, the type of test and the delay between reading and testing does not make much of a difference.

Experiment 2: Idea recognition tested immediately and after a delay

Inference verification may be inappropriate for demonstrating calibration because the domain of possible inferences is too broad. Verbatim recognition, it could be argued, may be inappropriate for demonstrating calibration because subjects do not represent the text in a verbatim manner (e.g., van Dijk and Kintsch, 1983). Instead, subjects may represent propositions, or ideas from the text. Thus a test of ideas, not requiring inferences and not requiring verbatim memory, might exhibit better calibration. We tested this conjecture in Experiment 2.

Method

Subjects. The subjects were 40 volunteers from introductory psychology classes at the University of Wisconsin, Madison. These subjects participated to fulfill a course requirement. Twenty subjects were randomly assigned to the two groups formed by the immediate versus delayed test variable.

Materials. The texts were identical to those used in the first experiment. In addition, we prepared a four-problem idea test for each text.

Each problem consisted of a close paraphrase of an idea in the text (see Appendix B for examples) and a distractor. The distractors were composed of words that were in the text, but they did not correspond to any idea in the text. The first problem (pair of ideas) always corresponded to an idea that was part of the text's central principle (used in the inference verification test).

Each confidence assessment form included the title of the appropriate text. Subjects were asked to circle a "number on the following scale to report how confident you are that you will be able to choose an idea from the text when given a choice between that idea and an idea not in the text." This statement was followed by the six-point scale. The corresponding four-problem idea recognition test appeared on the next page.

Procedure. Other than the use of the idea recognition test, the procedures were exactly as in Experiment 1 for the immediate and delayed conditions.

Results and Discussion

Insert Table 2

The results are presented in Table 2. All dependent measures were calculated twice, once using performance on the single idea recognition problem associated with the principle, and once using performance on all four problems.

Of the eight different measures of calibration, one, the product moment correlation in the immediate condition based on all four items, was significantly different from zero, t (16) = 2.34. None of the four differences between the immediate and the delayed conditions was significant, all $ps > .14$.

We draw three conclusions from these results. First, there is a hint that calibration for idea recognition is possible when the test is immediately after reading, but even this result might be a type 1 error. Second, imposing even a modest delay (about 20 minutes), drives calibration for idea recognition to zero. Third, the problem is not one of unreliability due to having a single test item. Even with four items, calibration in the delayed condition is not significantly different from zero, and the sign is negative.

Confidence Judgments Reflect Familiarity With the Text Domain, Not An Assessment of Comprehension

Calibration of comprehension is very poor. Empirically this means that there is little correlation between confidence judgments and performance on a test of comprehension. One explanation of this finding is that confidence judgments are random. An alternative is that subjects do make non-random assessments, but that the knowledge assessed is unrelated to the knowledge required for successful test performance.

Consider a domain familiarity strategy. When faced with a confidence assessment, the subject may use whatever information is provided on the confidence form (e.g., title, statement of principle) to assess familiarity with the domain of the text (rather than knowledge derived from the particular text). Domain familiarity then serves as the basis for confidence. Because general familiarity with a domain may not accurately predict performance on a test over a particular text, calibration may be low.

A domain familiarity strategy makes sense for two reasons. First, it might be much easier to assess familiarity with a domain than with a specific text. The distinction is akin to Reder's (1982) distinction between a relatively easy consistency judgment and a more difficult direct retrieval. Second, the strategy will lead to calibration when three conditions are satisfied: a) the range of texts samples multiple domains of knowledge, b) knowledge differs greatly across domains, and c) familiarity with the domains covaries with knowledge.

These conditions were met in Glenberg and Epstein (in press) in which music and physics students read texts in both music theory and physics. Knowledge differed across domains as indicated by differences in performance on the inference verification tests. Confidence also varied across domains. Finally, when considering texts across both domains, subjects, on the average, were calibrated ($\underline{G}$ = .24), even though they were not calibrated within a domain ($\underline{G}$ = .04).

Experiments 3-5 test predictions of the domain familiarity hypothesis. In Experiment 3 we demonstrate that familiarity with a domain predicts confidence assessments (but not performance). In addition, we demonstrate that the correlation between domain familiarity and confidence is greater than the correlation between recallability of the texts and confidence. In Experiments 4 and 5 we demonstrate that subjects can accurately judge familiarity of specific statements from a text, but that these judgments are not used in making confidence assessments. Apparently, domain familiarity, not familiarity with specific texts, controls confidence assessments.

Experiment 3: Domain familiarity predicts confidence ratings

It would seem a straightforward matter to determine if domain familiarity predicts confidence ratings: After reading each text, require the subject to judge domain familiarity and predict performance on a to-be-taken comprehension test, then correlate the two. We decided against this procedure because of the strong task demands. Namely, after rating familiarity, there is a strong demand to predict performance consonant with the familiarity rating. To eliminate these task demands, we had different subjects provide familiarity ratings and confidence ratings. Consequently, in this experiment, our conclusions only hold at the level of group data.

Three separate groups of subjects read the 15 texts. After reading, the subjects in group FCI provided a familiarity rating, a confidence rating, and performance on the inference test for each text. For this group, our interest was focused on the familiarity ratings. Subjects in group RCI recalled information from each text, provided a confidence rating for each, and took the

inference test for each text. For this group, our interest was in the recall. Finally, subjects in group CI provided confidence ratings and inference verification performance for each text.

For each text we computed the average familiarity rating from group FCI, the average recall of each text from group RCI, and the average confidence rating and inference test performance from group CI. On the assumption that familiarity with the text domains is relatively stable across our sample of subjects, the domain familiarity hypothesis makes the following predictions. First, the correlation between familiarity (from Group FCI) and confidence (from Group CI) should be substantial. Second, the correlation between recall (from group RCI) and confidence (from Group CI) should be less (because confidence is based on domain familiarity, not familiarity or recallability of a specific text). Third, familiarity should not correlate with performance on the inference verification task (from Group CI).

Method

Subjects. A total of 88 subjects from introductory psychology classes participated. There were 28 subjects in Group RCI, 30 subjects in Group FCI, and 30 subjects in Group CI.

Materials. The texts and the inference verification tests were the same as those used in Experiment 1. The Familiarity assessment form is reproduced in Appendix C. In short, it provides a direct quote of the principle from the text and requests a familiarity judgement from 1 to 6. The recall form (also reproduced in Appendix C) described the central principle and requested the subject to recall it exactly if possible. Finally, the confidence assessment form (similar to that used in Experiment 1) requested confidence in inference verification.

Procedure. All subjects were instructed to read the texts carefully, and that they would be tested using the inference verification procedure. After reading all of the texts, subjects were given special (written) instructions corresponding to the group to which they were assigned. As in the previous experiments, all materials were included in individual booklets and all phases of the experiment were self-paced.

Results and Discussion

The mean confidence ratings for groups FCI, RCI, and CI were 3.81, 4.29, and 4.13, respectively. These means were not significantly different, $F(2, 85) < 2.78$, $Mse = .63$. Mean performance on the inference test was between .68 and .70, and these means did not differ significantly between the groups, $F(2, 85) < 1$.

For each text we computed an average familiarity rating (from Group FCI), an average confidence rating (from Group CI), and an average performance (from Group CI). The recall data were treated as follows. Each subject's recall of each text was rated from 0 to 6 by two raters. The major criterion was the extent to which the recalled information corresponded to the central principle. Disagreements were resolved by discussion and by averaging. (A more objective measure of recall was also used; we simply counted the number of words in each recall protocol. The results using the two measures were very similar.)

Insert Table 3

Our first question is whether domain familiarity and recall will correlate with confidence ratings. The answer is provided by the data in Table 3. Note first that both familiarity and recall correlate with confidence. However, familiarity and recall also correlate with each other; partial correlations are needed to uncover the relationship between each variable and confidence with the contribution of the other variable partialled out. The partial correlations are also given in Table 3. The partial correlation of familiarity and confidence, .57, is highly significant, $t(12) = 2.43$. The partial correlation of recall and confidence, -.03, is not significant.

Insert Table 4

According to the domain familiarity hypothesis, the correlation between familiarity and inference performance should be low (that is why calibration is poor). The relevant data are provided in Table 4. Neither familiarity nor recall correlates highly with inference performance.

These results demonstrate that familiarity judgements are highly correlated with confidence judgments, and thereby support the claim that confidence is based on familiarity. Nonetheless, the procedure of Experiment 3 cannot distinguish between effects of domain familiarity and effects of familiarity with the particular texts. The inference that domain familiarity, not text familiarity, controls confidence is derived from the results of Experiments 4 and 5.

Experiments 4 and 5: Manipulating Statement Familiarity Does Not Affect Confidence

In Experiment 4 the FCI procedure was used. The major independent variable was the form of the statement used for familiarity assessment: Either a paraphrase or a verbatim restatement of the central principle was provided. The results demonstrate that this manipulation does affect familiarity with particular statements from the text. In Experiment 5 the CI procedure was used, and paraphrase or verbatim restatement of the principle was included on the confidence assessment form. We know from Experiment 4 that the paraphrase-verbatim manipulation affects familiarity of the statement. The question of interest is whether this manipulation will also affect confidence judgments. According to the domain familiarity hypothesis, subjects use information on the confidence assessment form to judge familiarity with the domain, not familiarity with a particular text or statement. Thus the hypothesis predicts no effect of paraphrasal on confidence judgments.

Because the materials and procedures were very similar, the two experiments are described together.

Method

Subjects. A total of 19 subjects participated in Experiment 4, and 20 subjects participated in Experiment 5.

Materials. For each text we wrote a new statement of each principle (see example in Appendix C). The new statement was written so that it could be directly substituted into the original text in place of the original principle. For each subject, approximately half of the texts contained the original principle, and half contained the new statement of the principle.

For each subject in Experiment 4, half of the familiarity assessment forms repeated the principle verbatim, and half presented a paraphrase of the principle (the version not in the text). Similarly, for each subject in Experiment 5, half of the confidence assessment forms repeated the principle verbatim, and half presented a paraphrase of the principle. Due to an error, the materials for one of the texts were transposed. This text was eliminated from all analyses.

Procedure. The procedures duplicated those used for Groups FCI and CI in Experiment 3.

Results and Discussion

Insert Table 5

The results are presented in Table 5. For the FCI group in Experiment 4, the .58 difference in familiarity ratings between the verbatim and paraphrase conditions was significant, $t(18) = 2.29$, $SE = .26$. This result demonstrates that the manipulation does affect familiarity with the particular statements.

The verbatim and paraphrase conditions did not differ significantly in regard to confidence, nor did they differ in performance on the inference tests.

For the CI group in Experiment 5, the -.08 difference in confidence ratings between the verbatim and the paraphrase conditions was not significant. Thus differences in familiarity with the particular statements on the confidence assessment form (demonstrated in Experiment 4) do not influence confidence assessments.

We began by proposing that calibration is poor because subjects do not assess the knowledge needed on comprehension tests, whether these are tests of inference verification, idea recognition, or verbatim recognition. The reason, according to the domain familiarity hypothesis, is that subjects assess familiarity with the general domain of the text, rather than familiarity with the specific statements on the confidence assessement form (that is, familiarity with the particular text).

Two forms of evidence (from these experiments) are consistent with the domain familiarity hypothesis. First, across subjects, familiarity with the

domains of the texts does significantly predict confidence ratings (Experiment 3). This prediction is not a trivial result of collapsing across subjects (and thereby increasing reliability of measures), because recallability of the texts did not significantly predict confidence ratings when the contribution of familiarity was partialled out. In other words, sometning peculiar to familiarity judgements is important.

Second, we asked the question, are confidence assessments controlled by familiarity with the domain of the texts or by familiarity with the particular statements used on the confidence assessment form. In Experiment 4 we demonstrated that we could easily manipulate familiarity with the specific statements. Nonetheless, this manipulation had no effect on confidence assessments in Experiment 5. These results demonstrate that familiarity with particular statements does not control confidence; the results are, by default, consistent with the domain familiarity hypothesis, although not conclusive.

Our confidence in the domain familiarity hypothesis is boosted by two analyses reported in Glenberg and Epstein (in press). In that study both music students and physics students read texts and took inference verification tests in both domains. For these students, knowledge in the domains varies greatly and familiarity covaries with that knowledge. Thus application of the domain familiarity strategy should result in across-domain calibration (not because subjects can accurately assess knowledge gained from a particular text, but because domain familiarity predicts performance across domains). Indeed these students were calibrated across domains.

The second analysis that demonstrated the operation of a domain familiarity strategy was as follows. For each subject Glenberg and Epstein (in press) determined (a) a single simulated confidence rating based on that subject's reported experience in music, and this simulated confidence rating was assigned to all music texts, and (b) a single simulated confidence rating based on that subject's reported experience in physics courses, and this simulated confidence rating was assigned to all physics texts. These confidence ratings simulate the operation of the domain familiarity strategy: Assign a confidence rating based on familiarity with the domain (not an assessment of knowledge gained from a particular text). Next, the simulated confidence ratings were used to compute simulated Gs for each subject. The mean simulated G was almost identical to the mean real G. Furthermore, the simulated Gs correlated .57 with the real Gs. Apparently, much of the predictive information in the confidence ratings is captured by application of a domain familiarity strategy.

It is not clear why subjects assess domain familiarity rather than familiarity with particular texts. As suggested before, it may be easier to assess domain familiarity than familiarity with particular texts. This may be especially so after reading many texts, as in these experiments.

Self-generated Feedback Can Be Used to Enhance Calibration

Apparently, calibration of comprehension is poor because subjects assess domain familiarity rather than knowledge gained from a particular text. Perhaps calibration can be improved if students can be taught to assess aspects of

knowledge more closely related to test performance than domain familiarity. In fact, the literature provides hints that this is the case; it appears that self-generated feedback from performance on a pre-test can be used to accurately predict future performance.

Consider studies of calibration reviewed by Lichtenstein, Fischhoff, and Phillips (1982). Subjects answered general knowledge questions and assessed the probability that the answers were correct. Generally, the correlation between performance and the probability assessments was quite high. We suspect this is so because subjects can use feedback obtained from answering the question (e.g., latency to answer the question, difficulty of any derivations, number of assumptions that had to be made) to assess the likelihood that the answer is correct.

Glenberg and Epstein (1985, in press) observed a similar type of calibration which they called performance calibration. After reading passages, subjects answered inference verification questions and judged the likelihood that their answers were correct. Although the subjects could not accurately predict performance, after taking the inference tests the subjects were accurate in judging the correctness of their answers. Again, self-generated feedback seems a likely source for this type of calibration.

Similar findings are reported in the domain of predicting memory performance. After studying a list of paired-associates (or sentences) once, subjects can relatively accurately predict cued recall (Lovelace, 1984). This predictive accuracy might well reflect subjects testing memory while making the predictions and using feedback from these self-tests to predict future test performance. In fact, both Lovelace (1984) and King, Zechmeister, and Shaughnessy (1980) have demonstrated that memory predictions improve after subjects are given an explicit test on the material.

Finally, data using the text calibration procedure are also consistent with the feedback hypothesis. In one of Glenberg and Epstein's (1985) experiments, subjects read texts and predicted performance on a second inference verification test after taking a first inference verification test (and predicting performance on the first test). Although predictions for the first test did not correlate with performance on the first test, predictions for the second test did significantly correlate with performance on the second test. Apparently, feedback gained from answering the first inference verification test can be used to predict performance on the second inference verification test.

In initial attempts to explicitly test the feedback hypothesis, subjects read the texts used in Experiments 1-5, and then answered a series of questions about each text. Some subjects had a pre-test consisting of two idea recognition problems (for each text). Next, subjects predicted performance on idea recognition post-test and then they took the post-test itself. Other subjects experienced the same sequence without the pre-test. Based on the feedback hypothesis, we predicted better calibration for subjects who took the pre-test.

In one of the intitial experiments the immediate test procedure (from Experiment 2) was used. The difference in calibration G between subjects who had the pre-test (n = 19) and those who did not (n = 19) was only .03. In a

second experiment the delayed test procedure was used. This time the difference between the Gs was .05 in the wrong direction.

At first glance, these results are incompatible with the feedback hypothesis. However, the fault may not be with the notion of feedback, but with implicit assumptions regarding the structure of the cognitive representation of the text. Consistent with current theorizing (e.g., van Dijk and Kintsch, 1983; Graesser, 1981), we assumed that the cognitive representation is abstract and highly interconnected. In this case, feedback based on testing one part of the representation should be valid for predicting performance based on a different (but connected) part of the representation. Unlike the experiments reported here, much of the research supporting the notion of interconnected representations has used narratives rather than exposition, relatively short and simple texts rather than naturalistic texts, and few texts before testing. In short, the assumption of abstract and interconnected representations may not hold for the experiments reported here.

Dropping the (implicit) assumption of interconnectedness, the feedback hypothesis can be modified and made more explicit: Feedback should be useful in predicting future performance only when the processes and knowledge that generate the feedback are relevant for the future test. That is, if the pre-test and post-test are independent (perhaps because they tap different knowledge), then feedback from the pre-test need not be predictive of post-test performance.

We used the data from the initial experiments to test this modified hypothesis. First, for each subject we computed the correlation between pre-test and post-test performance. Next, subjects were divided into groups on the basis of the sign of this correlation, and the average of individual subject Gs was computed for each group. The modified hypothesis predicts greater calibration for subjects whose pre-test performance correlates positively with post-test performance than for subjects for whom the correlation is not positive.

For the 19 subjects who had a pre-test in the immediate condition, 9 had positive correlations between the pre-test and the post-test, and 10 had negative correlations. The calibration Gs were .32 and .20, respectively. For the 20 subjects who had a pre-test in the delayed condition, 9 had positive correlations between the pre-test and the post-test, and 11 had negative correlations. The calibration Gs were .21 and -.02, respectively. In a third experiment using the delayed condition, 19 subjects took pre-tests (and rated latency to answer the pre-test questions). For the 10 subjects with postive pre-test-post-test correlations the average calibration G was .43, whereas for the 9 subjects with negative correlations the average calibration G was -.27. Thus, in all three initial experiments, the modified hypothesis was supported.

Experiments 6-8: Tests of the Modified Feedback Hypothesis and a Model

The modified feedback hypothesis is that feedback from a pre-test can be used to predict performance on a post-test to the extent that the processes and knowledge required on the post-test are similar to the processes that

generated the feedback. Experiments 6-8 tested two predictions generated from this hypothesis. The first prediction is that subjects will be calibrated on a post-test when the pre-test and the post-test use the same problems (unbeknownst to the subjects). This condition maximizes the similarity between the two tests and should maximize the predictive validity of the pre-test feedback. A second prediction is that a post-test consisting of problems unrelated to the pre-test to should produce little calibration.

A third, but more tentative, prediction is that a posttest consisting of problems that are different from but related to the pre-test should produce calibration intermediate between the same and unrelated post-tests. The prediction is tentative because it depends on our success in producing a related post-test. If the cognitive representation of the text is abstract and interconnected, then nominally related items may be closely connected in the representation and act much like the same items on the post-test. On the other hand, if the cognitive representation is (as our initial experiments led us to suspect) not closely connected, then problems that are nominally very similar may act as unrelated items.

To help explore the issue of degree of connectedness, as well as other issues raised by the data of Experiment 6, we developed a simple mathematical model of calibration based on feedback. The model is presented in the discussion of Experiment 6.

General Method for Experiments 6 - 8

The experiments reported in this section used similar procedures to test the modified feedback hypothesis. In all of the experiments subjects read 16 texts (15 were modified from the other experiments plus one additional). Subjects then received a series of 16 pre-tests and confidence assessments. For each text, the pre-test consisted of a single idea recognition problem with a confidence assessment on the same page (see Appendix D). The confidence assessment required a prediction as to performance on an idea-recognition post-test. Following the pre-tests and confidence assessments the subject received 16 post-tests (see Appendix D). Each post-test consisted of 3 idea recognition problems. The Same problem was identical to the problem used on the pre-test; the Related problem was a paraphrase of the Same problem; the Unrelated problem was from the text, but not closely related to the Same problem.

Ideas for the idea recognition tests were obtained using the following procedure. First, for each text, an idea (call it A) was identified. At some other point in the text we inserted a paraphrase of A (call it B). A second idea, relatively unrelated to A (call it C) was also identified, and a paraphrase of C (call it D) was also inserted into the text. The paraphrases were written to be intersubstitutable; that is, they could literally replace one another in the text without changing the meaning.

Each of the phrases A, B, C, and D served as old ideas on the idea recognition test. The distractors were constructed from content words that appeared in the passage, but the content words were reordered so as not to refer to any idea in the passage. Additionally, the distractor for A was a paraphrase of the distractor for B, and the distractor for C was a paraphrase of the

distractor for D (although these paraphrases were not particularly close, because the intersubstitutability criterion could not be applied).

For each text and for each subject an idea recognition problem was chosen for the pre-test (for example, the problem using idea A). The pre-test problem was counterbalanced for both texts and subjects. This problem was repeated on the post-test as the Same problem. The post-test also included the idea recognition problem using the paraphrase of the pretest idea (e.g., B). This was the Related problem. One of the two remaining ideas (e.g., C or D) was also included on the post-test as the Unrelated problem. Order of the three problems on the post-test was randomized. They were not identified to the subject as same, related, or unrelated.

Before reading the 16 texts subjects read two practice texts, took the pre-tests for the texts, and filled out the confidence assessments. The practice did not include the post-tests, just a blank piece of paper indicating that post-tests would be presented for the other 16 texts. The practice post-test was eliminated so that the subjects would not be forewarned (before filling out the confidence assessments) that some items would be repeated on the post-tests.

Subjects (Experiment 6). A total of 48 volunteers from introductory psychology classes at the University of Wisconsin served in the experiment to fulfill a course research requirement.

Results and Discussion (Experiment 6)

Two subjects were eliminated from the analyses because calibration measures could not be computed. This occurs when there is no variance in either the confidence ratings or the performance data. All data analyses were conducted using data from the remaining 46 subjects. The data of most interest are presented in Table 6. Each of the correlations was computed separately for each subject (based on the 16 texts). The means of the correlations are included in Table 6.

Insert Table 6

The calibration for the Same idea, r =.13, was significantly different from zero, $t(45) = 3.27$, SE = .04, as was the calibration for the Related idea r = .12, $t(45) = 3.39$, SE = .03. Calibration for the Unrelated idea was not significantly different from zero, r = .08, $t(46) = 1.87$, SE = .04.

Although there was some calibration in this experiment, it cannot be viewed as strong confirmation of the predictions from the modified feedback hypothesis. First, even calibration for the Same idea is very modest. Second, there is very little difference in calibration between the Same and the Related ideas, and not much of a difference between the calibrations for Same idea and the Unrelated idea. These failures of the modified feedback hypothesis cannot be because the initial conditions were not met: The pre-test is more closely related to the Same idea than to the Related or Unrelated items. Note that the correlation between performance on the pre-test and

performance on the Same idea is .56, whereas the correlations between the pre-test and the Related and Unrelated items were .30 and .00, respectively.

Nonetheless, there is some cause for worry about these data. Note that performance on the pre-test was low (62%). Additionally, the correlation between pre-test performance and confidence was low (.16). This latter datum may indicate that subjects cannot gain accurate feedback from the pre-test, or perhaps that feedback cannot be used when performance is so low.

Insert Figure 1

We devised a simple mathematical model to explore these issues. The major assumptions of the model are illustrated by the transition diagram in Figure 1. The model makes a distinction between knowledge and the belief that one is knowledgeable. Knowledge controls performance on the tests. It is acquired both from the text and from pre-experimental learning. Belief that one is knowledgeable controls both confidence ratings and consistency in responding from the pre-test to the post-test. In the absence of feedback, the major factor contributing to belief in knowledge is application of the domain familiarity hypothesis. When a pre-test is given, the major factor controlling belief is feedback from the pre-test.

For a particular problem, subjects with knowledge will always be correct, whereas those without knowledge will be correct with a probability equal to .5. When a problem is answered on the basis of knowledge, feedback from answering the problem will always lead to belief in knowledge. Therefore, subjects with knowledge will always believe that they have knowledge, and hence they will always use a high confidence rating. On the other hand, feedback from answering a problem will sometimes (with probability equal to b) lead subjects to believe that they have knowledge, when they do not. These subjects will also use a high confidence rating (but will sometimes be wrong on the test). Only subjects who do not believe that they have knowledge (with probability of $(1-k) \times (1-b)$) will use a low confidence rating.

Three other assumptions are needed. First, on Same idea problems, when the subject believes that he or she has knowledge, the subject's choice of alternatives will be the same as on the pre-test. When the subject believes that he or she is ignorant, the choices will be independent. Second, on Unrelated idea problems, the choice of alternatives is independent of the pre-test. Third, there is a probability v that a Related idea problem requires the same knowledge as the pre-test problem. Thus with probability v the Related idea problem is treated as a Same idea problem, and with probability $1-v$ the Related idea is treated as an Unrelated idea.

After estimating the three free parameters (k, b, v), the model can be used to derive the probabilities of various events such as the probability of high confidence correct choices and low confidence incorrect choices (see Appendix F). These probabilities can then be used to compute Phi coefficients that can be compared to the data. In addition, the value of the parameter v gives some indication of the connectedness of the cognitive representation. Given that the Related problem is a paraphrase of the pre-test, a high value of v is expected if the representation is abstract and interconnected.

We estimated values for the three parameters informally.[1] As a first approximation, we set $\underline{b}$ =.5, indicating that when a subject's knowledge is inadequate, the subject believes that knowledge is adequate half the time. Next, we chose a value for $\underline{k}$ that produced a reasonable prediction for the probability correct on the tests. Finally, $\underline{v}$ was chosen to predict the value of the P.R correlation exactly; thus in regard to the Related idea, the model is tested by the fit to the the C.R correlation (Related idea calibration). The predicted correlations are listed in Table 6.

Given the parameter estimation procedure, the pattern of predicted correlations is satisfyingly close to the data. Study of the model's structure and predictions revealed three other points. First, the maximum calibration is .71 when $\underline{k}$ = .99 and $\underline{b}$ = 0.0. Under the more reasonable assumption that $\underline{b}$ = .5, the maximum calibration is only .5. Thus our observed calibration of .13, although small, is not unreasonable.

Second, the correlation between pre-test performance and confidence (P.C correlation) is predicted to be equal to the correlation between confidence and Same idea performance (Same idea calibration). This prediction is made because the processes that generate performance and confidence (feedback) for the pre-test are exactly the same as the processes that generate performance for the Same idea on the post-test. Thus the low correlation between pre-test and confidence (see Table 6) must, according to the model, constrain Same idea calibration.

Third, the correlation between pre-test performance and confidence is a function of knowledge ($\underline{k}$): As knowledge increases the correlation increases. The increase is due to the elimination of low confidence correct responses due to guessing. Importantly, because of this relationship between level of knowledge and the P.C correlation, and because the P.C correlation constrains Same item calibration, when pre-test performance is low (as in the experiment), Same item calibration will also be low (as in the experiment).

These observations led us to perform another test of the modified feedback hypothesis (and the model based on it). The only change was to rewrite some of the idea recognition problems to boost performance on the pre-test. According to the model, increasing performance ($\underline{k}$) should increase the P.C correlation and increase Same idea calibration.

Method (Experiment 7)

Subjects. A total of 48 volunteers from introductory psychology classes at the University of Wisconsin served in the experiment to fulfill a course research requirement.

Materials and procedures. We rewrote those idea recognition problems associated with the lowest correct performance. Otherwise, the materials and procedures were identical to those used in Experiment 6.

Results and Discussion (Experiment 7)

Insert Table 7

A total of 38 subjects remained after eliminating those for whom no calibration measures could be computed due to lack of variance. The data for these 38 subjects are reported in Table 7. The correlations reported in the table are means of correlations computed separately for each subject.

Mean performance on the pre-test and post-test improved to about .78 correct. As predicted by the model, the average correlation between the pre-test and confidence also increased (compared to Experiment 6, Table 6), as did the average calibration for the Same idea.

In this experiment, the predictions of the modified feedback hypothesis are nicely supported. Statistical analyses are reported for the point-biserial correlations. The pattern of significant results is identical for analysis of G. First, calibration of the Same idea is significantly different from zero, $t(37)=7.53$ SE =.03. Also, judging from the size of the G coefficient, the effect is sizeable (for pairs of texts that differ in both confidence and performance, there is a .40 difference between the probability that the text with the greater confidence is correct and the probability that the text with the lower confidence is correct).

Second, calibration for the Same idea is significantly greater than calibration for the Related idea, $t(37)=4.14$, SE =.05. In fact, calibration for the Related idea is not significantly different from zero.

Third, the model successfully predicts other patterns in the data. Note that there does appear to be a close relation between the P•C correlation and the C•S correlation (Same idea calibration). The model also gives us some confidence in the claim that the low levels of calibration are not due to unreliability of the measures of confidence and performance. Note that the model assumes perfectly reliable measures of confidence and performance (when there is knowledge). Nonetheless, the model predicts low calibration.

From the perspective of research on text comprehension, these results are quite extraordinary in two ways. First, many theories of text comprehension propose that the result of reading is a representation composed of relatively abstract components such as propositions and macro-propostions (van Dijk and Kintsch, 1983), or schemata (Graesser, 1981). Our data imply that the representations used in these experiments are closely related to the surface structure (compare to Hayes-Roth and Thorndyke, 1978). This implication is based on the comparison of the Same idea and the Related idea. Remember, these ideas are intersubstitutable paraphrases of one another, and both of the ideas occurred verbatim in the text. Nonetheless, the average correlation between performance on the Same idea and the Related idea was only .17 (.30 in Experiment 6). Also, calibration of the Related idea was only .07. Apparently, the Same and Related items are not retrieving the same information from memory. In terms of the mathematical model, the probability that the two problems contact the same knowledge is only .19 (v).

Second, most theories of text comprehension propose that the representation of text is well-organized and connected. Our data imply that the representation is not highly connected. This implication is based on the very poor calibration in all but the Same idea condition. If the representation was highly connected, then feedback from the pre-test should have predicted of performance on any other idea from the text. As the data demonstrate, however, this was not so.

Perhaps a surface-based, unconnected representation is characteristic of reading many short, unfamiliar, expository texts in close contiguity. Alternatively, the data might simply reflect a surface-structure strategy. Note that all of the old ideas in the idea recognition problems consisted of strings of words that were contiguous in the text, whereas the distractor ideas consisted of strings of words that were not contiguous in the text. Now suppose that text is represented at multiple levels (van Dijk and Kintsch, 1983; Johnson-Laird, 1983) with one level being close to the surface structure and other levels being more abstract. Because old and new ideas can be discriminated simply by comparison to the surface representation, subjects may have adopted the strategy of consulting only the surface representation. This alternative was tested in the next experiment.

Experiment 8 was identical to Experiments 6 and 7, except for one substantive change. In the previous experiments, the old ideas used in the idea recognition problems were taken verbatim from the texts. In Experiment 8, the old ideas were paraphrases of ideas used in the texts. This change precludes the use of a surface matching strategy to discriminate between old and new ideas.

Method (Experiment 8)

Subjects. The 48 subjects were volunteers from the same source as used previously.

Materials and procedures. For each text we used ideas A, B, C, and D, and we wrote paraphrases of each of these ideas (A', B', C', and D'). The paraphrases used different content words, and in no case did the paraphrase appear verbatim in the text. We also wrote paraphrases for the original distractors. These distractor paraphrases used content words that did not appear in the text. Examples appear in Appendix E.

For half the subjects, ideas A, B, C, and D appeared in the text and ideas A', B', C', and D' were used in the pre-test and post-test. For the remaining subjects the roles were reversed. The old ideas were always paired with the newly written distractors.

Other details of the design and procedure were identical to Experiment 7, except for one change in the instructions. Subjects were forewarned that the idea recognition problems would consist of paraphrases of ideas presented in the texts.

Results and Discussion (Experiment 8)

Insert Table 8

A total of 37 subjects remained after eliminating those for whom no calibration could be computed. The data for the 37 subjects are in Table 8.

Proportion correct was at a reasonable level, .73, so that we may expect to see calibration. Indeed, calibration for the Same idea was significantly greater than zero, $t(36) = 6.32$, SE = .04. Related idea calibration, although low, was also significantly different from zero, $t(36) = 2.33$, SE = .05.

The critical question is whether there is a significant difference between Same idea and Related idea calibration. Indeed, the difference was significant, $t(36) = 2.82$, SE = .06. Thus the difference between these two calibrations observed in the previous experiment cannot be attributed solely to the application of a surface strategy.

Ignoring for a moment the unrelated idea calibration, the model does a credible job of predicting the pattern of the data. Note that the P.C correlation is again very similar to the C.S correlation (Same idea calibration). The model also successfully predicts the relationship between the P.R correlation (determined in part by the parameter v) and the C.R correlation (also determined in part by v).

One surprising result is the level of Unrelated idea calibration. In the previous two experiments it was not significantly different from zero. In this experiment it was significant, $t(36)=4.06$, SE =.04. Based on the following reasoning, we believe that this is probably a type 1 error. First, in neither of the previous experiments was the Unrelated idea calibration significantly greater than zero: It was not significant in Experiment 6 which had more power than Experiment 8; and it was not significant in Experiment 7 (with comparable power) which had a higher proportion correct, and so presented a better opportunity for unrelated idea calibration to reveal itself. Second, it is difficult to imagine circumstances in which Unrelated idea calibration should be greater than Related idea calibration. Third, the model, which does a credible job of predicting calibration in the previous experiments and in this experiment, predicts low Unrelated idea calibration.

In sum, we draw two conclusions from the results of this series of experiments. First, given appropriate feedback (e.g., from a pre-test), calibration can be significant, and judging from the Same idea G, it can be considerable. Second, predictions based on feedback have a tightly circumscribed domain; that is, there is little transfer to Related idea problems. Judging from Experiment 8, the failure for predictions to transfer to the Related item is not due to application of a surface matching strategy.

General Discussion

To review, we have made three major claims. First, low calibration of comprehension is a general problem, not one confined to a particular mode of testing. Experiments 1 and 2 demonstrated that calibration is low when testing is by inference verification, verbatim recognition, or idea recognition. Also, calibration is low when testing is immediately after reading or after a modest delay.

Second, one cause of poor calibration is that people tend to assess general familiarity with a text domain, rather than knowledge gained from a particular text (or even familiarity with that particular text). Experiment 3 demonstrated (over subjects) that familiarity judgements predicted confidence ratings, but not inference verification performance. Once the contribution of familiarity was partialled out, recall did not predict confidence judgements. Experiments 4 and 5 demonstrated that familiarity with a specific sentence could be manipulated, but that that form of familiarity did not affect confidence judgements. Hence the conclusion that domain familiarity is a major determiner of confidence in comprehension.

This finding helps to account for the belief that self-assessments of knowledge are accurate. It is likely that performance is high in domains of high familiarity and lower in domains of low familiarity. Thus across domains of widely different familiarities, judgments of knowledge are likely to predict performance (Glenberg & Epstein, in press). Nonetheless, because it is so difficult to assess knowledge gained from a particular text, calibration of comprehension is generally low.

Third, calibration of comprehension can be improved by providing feedback in the form of a pre-test. This feedback is only useful, however, when the pre-test is very closely related to criterion (post-test) performance. This finding provides an empirical bridge between our work on calibration of comprehension and the work demonstrating good calibration in general knowledge tasks (Lichtenstein et al., 1982) and memory tasks (King, Zechmeister, and Shaughnessy, 1980; Lovelace, 1984). Even the finding that calibration is circumscribed can be related to the general knowledge and memory work. It seems likely that a subject's judgement regarding memorability of one of a list of items would not predict memorability of other items from that list. Our results from Experiments 6-8 are similar. After an overt pretest, confidence judgements accurately predict future performance on that same item; the confidence judgements predict less well future performance on related items.

The remainder of this discussion describes two implications of these findings.

The clearest implication is in regard to the accuracy of self-assessments of comprehension. If they are to be useful predictors of future performance, then the assessments should be based on feedback from a task similar to the criterion task. Judgements based on undifferentiated feelings of familiarity, although subjectively compelling, are not predictive of performance requiring knowledge from a particular text.

Our finding that feedback is only predictive when the same items are tested

on the pre-test and the post-test appears to put severe constraints on the usefulness of a pre-test. However, generalizing a step or two beyond the data changes this appearance. First, the constraints may apply only when the cognitive representation of the text is relatively unconnected. Increasing the connectedness of the representation (perhaps by using advance organizers, or other signals as to text organization), may enhance the generalizability of the feedback.

Second, according to the modified feedback hypothesis, feedback is useful in predicting future performance when the processes and knowledge that generate the feedback are relevant on the criterion task. Feedback from the idea recognition pre-test probably takes the form of a feeling of familiarity with the specific idea presented on the pre-test, or perhaps an estimate of the difficulty of performing the discrimination. There is little in this feedback that would be diagnostic regarding performance on a different idea recognition problem.

In contrast, consider the sort of feedback that can be gained from an inference verification pre-test. The inference verification task requires a number of processes, including retrieval of a representation of the central principle of the text (or perhaps constructing it out of whatever can be retrieved), and attempting to use the principle to derive the inference. Feedback from this sort of pre-test could indicate the difficulty of retrieving or constructing the principle, and this sort of feedback should be predictive of performance on any inference verification problem that requires retrieval of the same principle, not just the exact same inference verification problem.

This speculation is supported by data from an earlier experiment (Glenberg and Epstein, 1985, Experiment 3). In that experiment subjects were given two inference verification problems for each text. The two inference verification problems were not the same, but they were related in that both required knowledge of the same principle. The problems were separated by two confidence assessments (confidence that the answer to the last problem was correct, and confidence that the answer to the next problem would be correct). The first inference verification problem can be viewed as a pre-test, and the second can be viewed as a Related item post-test. Direct application of the results of Experiments 7 and 8 would lead to the prediction of poor Related item calibration. Alternatively, the modified feedback hypothesis suggests that the important factor is that the feedback is generated by processes and knowledge relevant on the post-test (such as retrieval of the principle). In this case, we expect a level of calibration on the Related inference verification post-test comparable to the level of calibration for the Same idea post-test in Experiments 7 and 8.

The data from the experiment were quite clear. Subjects were significantly calibrated on the post-test, having an average Pearson correlation of .19, which is in the same ballpark as our Same idea calibration in Experiments 6-8. Thus there is some support for the claim that what is important is not the overt similarity between the pre-test and the post-test, but the predictive utility of pre-test feedback--that is, the extent to which pre-test knowledge and processes are also used on the post-test.

Given this interpretation, the modified feedback hypothesis should be useful for designing informative pre-tests. The criterion for producing

informative feedback is that the processes and structures tapped by the pre-test should be the same as those required by the post-test. For example, an instructor might help a student prepare for an essay exam by providing the student with practice essay questions that require access to the same knowledge required on the exam. Feedback from attempting to answer the practice essays should be useful for predicting performance on the exam. For another example, consider the sort of pre-test that might be effective in a more performance-oriented domain (e.g., sailing). Since the criterion test involves procedural as well as declarative knowledge, a pre-test that only taps declarative knowledge is unlikely to provide useful feedback.

The second implication of these results is for theories of text comprehension. As noted earlier, many theories propose that the result of comprehension is an abstract integrated structure. Our data are at odds with this proposal.

Apparently, the structure tapped in these experiments was not very abstract, and not very integrated. Consider the results for Experiments 6-8. The mean correlation between pre-test performance and performance on the Related idea varied from .12 to .30. Remember, in these experiments the Related idea was an intersubstitutable paraphrase for the pre-test idea, and yet judging from these correlations the two ideas seem to have separate representations. In the model, the parameter v is the probability that the Related idea contacts the same knowledge as the pre-test idea. Estimates of this parameter range from .19 to .51. The highest value is from Experiment 6 in which performance was so low that the parameter estimates are probably unstable. Even so, the parameter estimates indicate that paraphrases are generally unlikely to contact the same representation.

What are we to make of this? One alternative is that these results are anomalous, perhaps due to the requirement that subjects had to read and remember so much. We cannot rule out this possibility. Nontheless, the time and effort required to read our texts (30-40 minutes) is probably not much different from the time and effort required to read a chapter in an introductory textbook. Thus at least some of the task demands are representative of real situations, and we see no reason to doubt that our results are representative.

Another alternative is that each text was represented by a collection of locally coherent structures, each based on one or two sentences, and each incorporating specific lexical items (Hayes-Roth & Thorndyke, 1979). Such a structure is likely to be constructed jointly from the subject's knowledge of the things (processes, events, objects) described by the text and the constraints inherent in the syntactic and semantic relations specified by the sentences. Rather than being abstract, each local structure may represent a specific situation (interpretation of the text) with which the subject is familiar, or when the material is unfamiliar, the particular words used in the test. Thus even intersubstitutable paraphrases (which use different content words) may result in different local structures. The representation of the text need not be tightly connected. Instead it may consist of a series of partially overlapping local structures each constructed to conform to the new constraints imposed by additional reading of the text.

We began by noting that readers are poorly calibrated. In conclusion, we note that we now have an understanding of why calibration is poor--the

misapplication of a domain familiarity strategy--and suggestions for improving calibration. Foremost is the suggestion of obtaining feedback on a pre-test that requires the same processes and structures as the criterion test. More tentatively, calibration based on a pre-test may be increased by developing an interconnected representation of the text.

References

Epstein, W., Glenberg, A. M., & Bradley, M. M. (1984). Coactivation and comprehension: Contribution of text variables to the illusion of knowing. Memory & Cognition, 12, 355-360.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 702-718.

Glenberg, A. M., & Epstein, W. (in press). Inexpert calibration. Memory & Cognition.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. Memory & Cognition, 10, 597-602.

Graesser, A. C. (1981). Prose comprehension beyond the word. New York: Springer-Verlag.

Hayes-Roth, B., & Thorndyke, P. W. (1979). Integration of knowledge from text. Journal of Verbal Learning and Verbal Behavior, 18, 91-108.

Johnson-Laird, P. N. (1983). Mental models. Cambridge, MA: Harvard University Press.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. American Journal of Psychology, 93, 329-343.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under certainty: Heuristics and biases. New York: Cambridge University Press.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 756-766.

Maki, R., & Berry, S. (1984). Metacomprehension of text material. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 663-679.

Maki, R. N., & Monson, N. S. (1985). Massed versus distributed study of text: Effects on test performance and predictions of test performance. Paper presented at the meetings of MPA, Chicago.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 288-294.

Nelson, T. D. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95, 109-123.

Nelson, T. D., Leonesio, R. J., Landwehr, R. S., & Norens, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing vs. the normative feeling of knowing vs. base-rate item difficulty. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 279-287.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. Psychological Review, 89, 250-280.

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

Author note

* Thomas Sanocki is now at the University of South Forida.

This research was sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under contract No. N0014-K-0644, Contract Authority Identification Number, NR 702-012. Requests for reprints should be sent to Arthur Glenberg, Department of Psychology, W.J. Brogden Psychology Building, University of Wisconsin, Madison, Wisconsin 53706.

Footnote

[1]Because the quantities estimated are non-independent, formal measures of goodness of fit are inappropriate. Nonetheless, as a check on our informal parameter estimates we used an iterative curve fitting program to find parameters that minimized the sum of the squared deviations between the observed and the predicted values. The "best fitting" parameters were all close to those reported in the text.

Table 1

Data from Experiment 1 (standard deviations in parentheses)

	Verbatim Recognition		Inference Verification	
Dependent Variable	Immediate[a]	Delayed[b]	Immediate[c]	Delayed[b]
Confidence	4.65 (.69)	4.22 (.70)	4.84 (.65)	5.04 (.73)
Proportion Correct	.87 (.10)	.71 (.15)	.79 (.14)	.81 (.12)
Calibration r	-.09 (.29)	-.04 (.23)	.04 (.29)	.06 (.29)
Calibration G	-.13 (.69)	-.11 (.47)	.14 (.65)	.09 (.72)

[a] n = 20 for confidence and proportion correct, n = 15 for calibration

[b] n = 20 for confidence and proportion correct, n = 18 for calibration

[c] n = 20 for all measures

Table 2

Data from Experiment 2 (standard deviations in parentheses)

Dependent Variable	Immediate[a]	Delayed[b]
Confidence	4.27 (.99)	3.82 (.66)
Proportion Correct (principle)	.68 (.16)	.74 (.12)
Proportion Correct (total)	.77 (.48)	.78 (.44)
Calibration r_{pb} (principle)	.07 (.30)	-.10 (.34)
Calibration r_{pb} (total)	.21 (.37)	-.03 (.29)
Calibration G (principle)	.12 (.56)	-.26 (.58)
Calibration G (total)	.22 (.54)	-.13 (.34)

[a] n = 20 for confidence and proportion correct, n = 16 for principle calibration, n = 17 for total calibration

[b] n = 18 for confidence and proportion correct, n = 17 for calibration measures

Table 3

Correlations and Partial Correlations (in parentheses) with Confidence

	Familiarity	Recall
Recall	.63	
Confidence	.66 (.57)	.40 (-.03)

Table 4

Correlations and Partial Correlations (in parentheses) with Inference Verification Performance

	Familiarity	Recall
Recall	.63	
Inference	.04 (-.07)	.15 (.16)

Table 5

Data for Experiments 4 and 5

	Familiarity	Confidence	Inference
Experiment 4 (FCI)			
Verbatim	5.17	5.19	.74
Paraphrase	4.59	5.02	.76
Experiment 5 (CI)			
Verbatim	-	4.87	.80
Paraphrase	-	4.95	.78

Note: The familiarity and confidence data are mean ratings on a scale of 1-6. The inference data are mean proportions correct.

Table 6

Data from Experiment 6 and Predictions from the Model

	Observed	Predicted
Pre-test Performance (P)	.62	.60
Confidence (C)	4.01	-
Same-item Performance (S)	.64	.60
Related-item Performance (R)	.65	.60
Unrelated-item Performance (U)	.68	.60

Pretest Correlations

	Observed r	Predicted r
P•C	.16	.16
P•S	.56	.58
P•R	[illegible]	.30
P•U	[illegible]	.00

Calibration Correlations

	Observed r (G)	Predicted r
C•S	.13 (.10)	.16
C•R	.12 (.14)	.08
C•U	.08 (.04)	.00

Note: Predictions are based on the following parameter values:

$k = .19$, $b = .50$, $v = .51$.

Table 7

Data from Experiment 7 and Predictions from the Model

	Observed	Predicted
Pre-test Performance (P)	.78	.73
Confidence (C)	3.76	-
Same-item Performance (S)	.78	.73
Related-item Performance (R)	.79	.73
Unrelated-item Performance (U)	.78	.73

Pretest Correlations

	Observed r	Predicted r
P·C	.31	.31
P·S	.55	.65
P·R	.12	.12
P·U	.04	.00

Calibration Correlations

	Observed r (G)	Predicted r
C·S	.26 (.40)	.31
C·R	.07 (.08)	.06
C·U	.04 (.07)	.00

Note: Predictions are based on the following parameter values:

k = .45, b = .50, v = .19.

Table 8

Data from Experiment 8 and Predictions from the Model

	Observed	Predicted
Pre-test Performance (P)	.73	.68
Confidence (C)	3.95	-
Same-item Performance (S)	.72	.68
Related-item Performance (R)	.74	.68
Unrelated-item Performance (U)	.72	.68

Pretest Correlations

	Observed r	Predicted r
P•C	.25	.26
P•S	.57	.63
P•R	.22	.22
P•U	.11	.00

Calibration Correlations

	Observed r (G)	Predicted r
C•S	.27 (.35)	.26
C•R	.11 (.12)	.09
C•U	.18 (.22)	.00

Note: Predictions are based on the following parameter values:

k = .35, b = .50, v = .35.

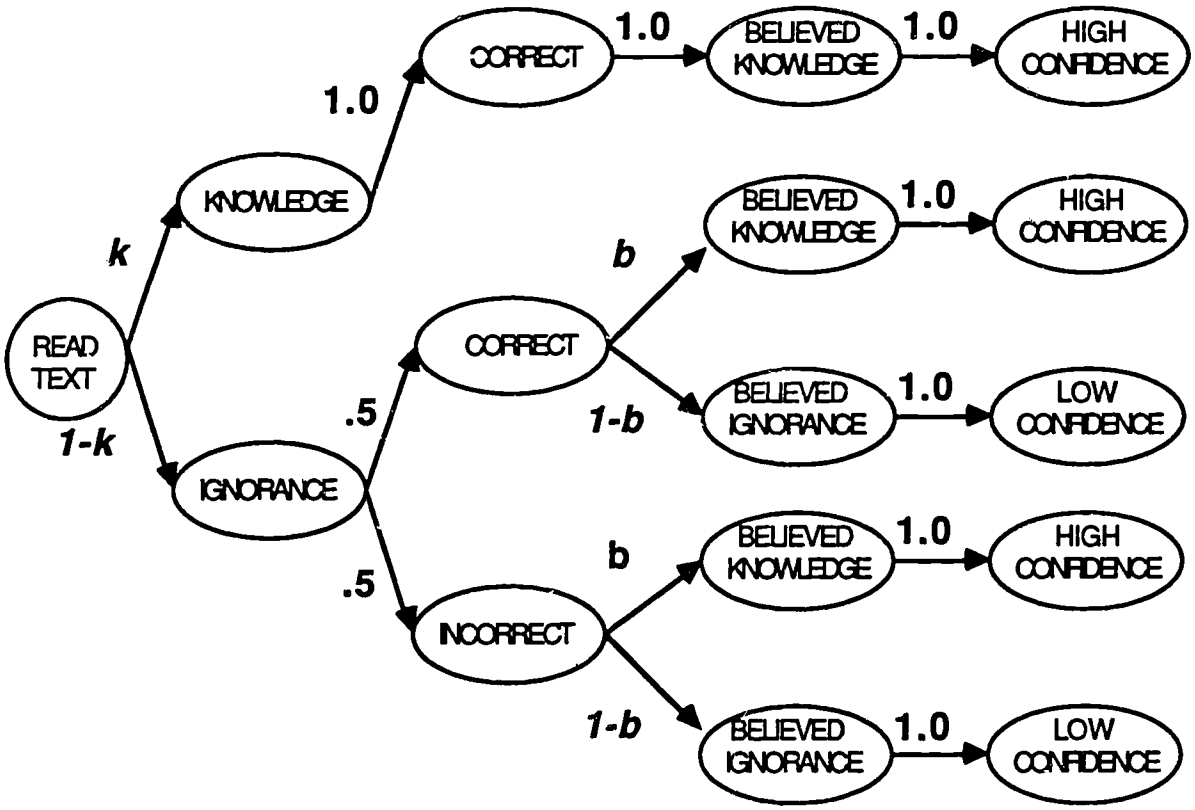


Figure 1. Model for calibration.

Appendix A:

Sample Materials for Experiment 1

Text:

Control of Eating by Blood Sugar

For most animals hunger is virtually permanent as a result of difficulties in obtaining food; these animals eat whenever food is obtainable. However, for mammals with a plentiful food supply, hunger and consumption of food is regulated by the hunger and satiety eating control centers in the brain. These centers are sensitive to the level of glucose circulating in the blood. Increases in blood glucose stimulate the satiety center and thereby reduce eating; decreases in blood glucose stimulate the hunger center and thereby induce eating. Shortly after a meal, when the concentration of glucose in the blood is high, the satiety center signals a state of fullness prompting the animal to refuse food. Many hours after a meal, when the concentration of glucose in the blood is low, the hunger center responds and the animal is prompted to eat.

Confidence assessment for verbatim recognition test:

Control of Eating by Blood Sugar

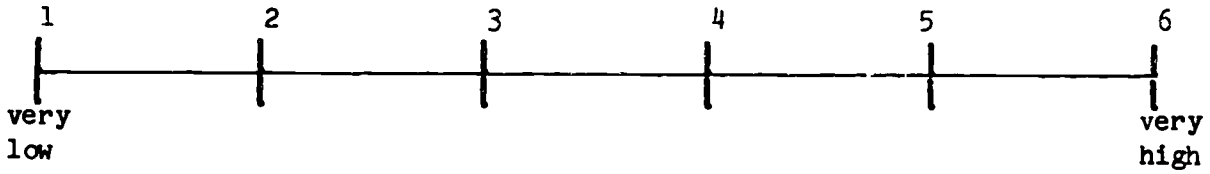
Circle a single number on the following scale to report how confident you are that you will be able to choose a verbatim (word for word) sentence from the text when given a choice between a verbatim sentence and a paraphrase (restatement of the sentence).

Confidence assessment for inference verification tests:

Control of Eating by Blood Sugar

One of the central points of this text dealt with the topic listed below. Circle a single number on the following scale to report your confidence that you are able to use what you have learned in this text to draw correct inferences using that point.

Increases in blood glucose stimulate the satiety center and thereby reduce eating; decreases in blood glucose stimulate the hunger center and thereby induce eating.



Verbatim recognition test:

Control of Eating by Blood Sugar

1. Rising levels of glucose in the blood activate the satiety center and thereby reduce eating; lowering of blood glucose activates the hunger center and arouses eating.

2. Increases in blood glucose stimulate the satiety center and thereby reduce eating; decreases in blood glucose stimulate the hunger center and thereby induce eating.

Inference verification test (true version):

Control of Eating by Blood Sugar

Inference: Intravenous injections of insulin lower glucose concentrations in the blood. An intravenous injection of insulin will cause a mammal who is sated to eat more.

T F

Inference verification test (false version):

Control of Eating by Blood Sugar

Inference: Intravenous injections of insulin lower glucose concentrations in the blood. An intravenous injection of insulin will cause a mammal who is hungry to refuse food.

T F

Appendix B:

Sample Idea Recognition Problems from Experiment 2

Control of Eating by Blood Sugar

1. satiety center
2. glucose control center

1. glucose blood level
2. concentration of blood cells

1. regulation of hunger
2. difficulties in blood circulation

1. glucose abnormalities
2. state of fullness

Appendix C:

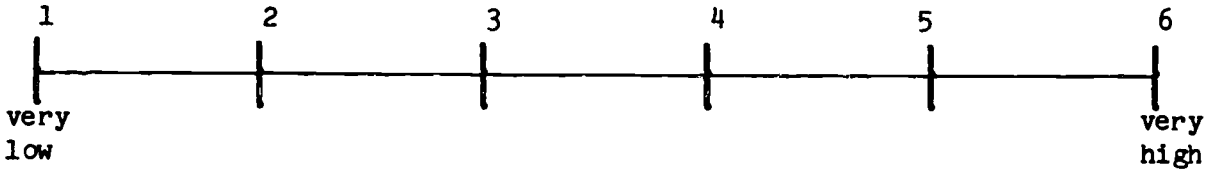
Sample Materials from Experiment 3

Familiarity assessment (Group FCI):

Control of Eating by Blood Sugar

Circle a single number on the following scale to indicate how familiar the following statement from the above passage appears to you.

"Increases in blood glucose stimulate the satiety center and thereby reduce eating; decreases in blood glucose stimulate the hunger center and thereby induce eating."



Recall probe (Group RCI):

Control of Eating by Blood Sugar

One of the central points of this text dealt with the topic listed below. Please try to write down that point in one or two sentences, exactly as stated in the text if possible.

The mechanisms by which increases and decreases in blood glucose affect eating

Appendix D:

Sample Materials for Experiments 6 and 7

Text:

Control of Eating by Blood Sugar

For most animals hunger is virtually permanent as a result of difficulties in obtaining food. These animals cannot count on a regular supply of food so they eat whenever food is obtainable. However, for animals with a plentiful food supply, hunger and consumption of food is regulated by the hunger and satiety eating control centers (idea A) in the brain. These centers are tiny regions in the hypothalamus that contain receptors (cells) that respond to biochemicals in blood. In particular, these centers react to variations in the density of blood sugar (idea C) (glucose) circulating in the blood. Rising levels of glucose in the blood activate the satiety center and thereby reduce eating; lowering of blood glucose activates the hunger center and arouses eating. Shortly after a meal, the increasing concentration of glucose (idea D) causes the satiety center to signal a state of fullness prompting the animal to refuse food. Many hours after a meal, when the concentration of glucose is low, the hunger center responds and the animal is prompted to eat. Activation of these two feeding regulation sites (idea B) allows the animal to avoid the problems of over- and undereating.

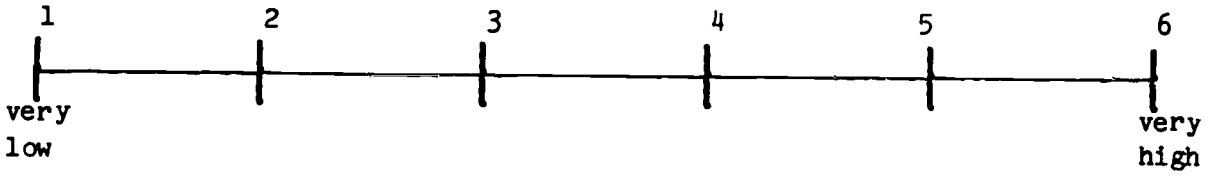
Pre-test:

Control of Eating by Blood Sugar

1. permanent fullness
2. density of blood sugar

Confidence assessment:

Consider your experience in choosing between the pair of ideas on the immediately prior test. Use this experience to estimate how confident you are that you will be able to choose another idea from the text, when given a choice between that idea and an idea not in the text.



Post-test:

Control of Eating by Blood Sugar

(Related)
1. regular satiety
2. concentration of glucose

(Same)
1. density of blood sugar
2. permanent fullness

(Unrelated)
1. regulation problems
2. eating control centers

Appendix E:

Sample Item Recognition Test for Experiment 8

Control of Eating by Blood Sugar

(Unrelated)	1. amount of blood sugar 2. constant satiation
(Same)	1. difficulties of regulation 2. food intake control sites
(Related)	1. complications for control processes 2. consumption regulation centers

Appendix F:

Illustrative Derivations from the Model

$$p(\text{correct}) = p(\text{knowledge}) + p(\text{ignorance}) \times .5$$

$$= \underline{k} + (1 - \underline{k}) \times .5$$

Given a 2 x 2 table such as that below, the phi correlation is equal to

$$r_{\phi} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

A	B	A+B
C	D	C+D
A+C	B+D	1.0

Finding the four probabilities corresponding to A, B, C, and D allow computation of the relevant correlations. For example, for the pre-test performance, confidence correlation (P·C):

$$p(\text{correct \& high confidence}) = p(\text{knowledge}) + p(\text{ignorance}) \times .5 \times p(\text{believed knowledge})$$

$$= \underline{k} + (1-\underline{k}) \times .5 \times \underline{b}$$

$$p(\text{correct \& low confidence}) = p(\text{ignorance}) \times .5 \times p(\text{believed ignorance})$$

$$= (1-\underline{k}) \times .5 \times (1-\underline{b})$$

$$p(\text{incorrect \& high confidence}) = p(\text{ignorance}) \times .5 \times p(\text{believed knowledge})$$

$$= (1-\underline{k}) \times .5 \times \underline{b}$$

$$p(\text{incorrect \& low confidence}) = p(\text{ignorance}) \times .5 \times p(\text{believed ignorance})$$

$$= (1-\underline{k}) \times .5 \times (1-\underline{b})$$

The 2 x 2 matrix for the P•C and C•S correlations is:

	Correct (on P)	Incorrect (on P)	
high confidence	$k + (1-k) \times .5 \times b$	$(1-k) \times .5 \times b$	$k + (1-k) \times b$
low confidence	$(1-k) \times .5 \times (1-b)$	$(1-k) \times .5 \times (1-b)$	$(1-k)\ (1-b)$
	$k + (1-k) \times .5$	$(1-k) \times .5$	1.0

The 2 x 2 matrix for the P•S correlation is:

	Correct (on P)	Incorrect (on P)	
correct (on S)	$k + (1-k) \times (1+b) \times .25$	$(1-k) \times (1-b) \times .25$	$k + (1-k) \times .5$
incorrect (on S)	$(1-k) \times (1-b) \times .25$	$(1-k) \times (1+b) \times .25$	$(1-k) \times .5$
	$k + (1-k) \times .5$	$(1-k) \times .5$	1.0

The 2 x 2 matrix for the P•U correlation is:

	Correct (on P)	Incorrect (on P)	
correct (on U)	$[k + (1-k) \times .5]^2$	$[(1-k) \times .5][k+(1-k) \times .5]$	$k+(1-k) \times .5$
incorrect (on U)	$[k+(1-k) \times .5][(1-k) \times .5]$	$[1-k \times .5]^2$	$(1-k) \times .5$
	$k + (1-k) \times .5$	$(1-k) \times .5$	1.0

The formula in each cell of the P•R matrix is v times the corresponding cell in the P•S matrix plus $(1-v)$ times the corresponding cell in the P•U matrix.

The 2 x 2 matrix for the C•U correlation is:

	Correct (on U)	Incorrect (on U)	
high confidence	[k+(1-k)xb] [k+(1-k)x.5]	[k+(1-k)xb] [(1-k)x.5]	k+(1-k)xb
low confidence	[(1-k)x(1-b)] [k+(1-k)x.5]	[(1-k)x(1-b)] [(1-k)x.5]	(1-k)x(1-b)
	k + (1-k) x .5	(1-k) x .5	1.0

The formula in each cell of the C•R matrix is v times the corresponding cell in the C•S matrix plus (1-v) times the corresponding cell in the C•U matrix.