Memory & Cognition
1987, 15(1), 84-93

Inexpert calibration of comprehension

ARTHUR M. GLENBERG and WILLIAM EPSTEIN
University of Wisconsin, Madison, Wisconsin

Students with a wide range of course work in physics or music theory read expositions in both domains. After reading 16 texts, each student provided a judgment of confidence in his/her ability to verify inferences based on the central principles of the texts. The primary dependent variable was calibration of comprehension, the degree of association between confidence and performance on the inference test. Two results of most interest were that (1) expertise in a domain was inversely related to calibration and (2) subjects were well calibrated across domains. Both of these results can be accommodated by a self-classification strategy: Confidence judgments are based on self-classification as expert or nonexpert in the domain of the text, rather than on an assessment of the degree to which the text was comprehended. Because self-classifications are not well differentiated within a domain, application of the strategy by experts produces poor calibration within a domain. Nonetheless, because self-classification is generally consistent with performance across domains, application of the strategy produces calibration across domains.

A reader's self-assessment of comprehension often has significant consequences for the reader's action. When reading under time constraints, the reader may believe that comprehension has been achieved, which encourages the reader to terminate further processing of the text. When reading in preparation for testing, the reader may believe that comprehension has been attained, which leads the reader to declare his or her readiness for testing. Given these and other implications for action, it is sensible to inquire whether readers' beliefs are regularly valid. By measuring the relationship between the readers' self-assessments of confidence in comprehension (strength of belief) and performance on a test of comprehension, we have repeatedly found that readers' beliefs typically are off the mark. Readers are very poorly *calibrated*: confidence in comprehension (belief) does not predict performance.

Glenberg and Epstein (1985) measured calibration by having subjects read 15 short expositions on a variety of topics. Subjects also provided an assessment of their confidence in their ability to use a principle from the text (provided at the time of the confidence assessment) to judge whether or not an inference was correct. Finally, subjects attempted to decide whether an inference using the principle was or was not valid. One measure of calibration of comprehension is the point-biserial correlation between the confidence assessments and performance on the inference test. In none of three experiments reported by Glenberg and Epstein was this correlation significantly different from zero.

In a similar study, Maki and Berry (1984) had subjects read a chapter from an introductory psychology text, rate comprehension, and then take a comprehension test. They also found generally poor calibration. For example, on an immediate test over the first half-chapter, calibration was low but significantly different from zero. On a test over the second half-chapter, calibration was essentially zero.

In subsequent experiments (Glenberg & Epstein, 1986) deploying a variety of performance measures and a diverse set of measures of calibration, the finding of zero or marginal calibration has recurred. This result is disconcerting because it appears to identify an important obstacle in learning from text. The result also does not conform to our personal experience. In our experience in learning from text, calibration of comprehension seems reasonably good.

After more detailed scrutiny of our experience, our initial impression that, in general, we were calibrated had to be qualified. Our impression may have been much affected by the availability heuristic. In assessing the degree of calibration that we exhibited, we relied heavily on the most readily available texts, and as a matter of course, these texts were in our personal domains of expertise. By contrast, in our experiments, the texts were by design a varied set that probably touched only peripherally on readers' special fields of competence. These considerations led to the current experiment to test the relationship between calibration and expertise.

Everyday observation suggests that experts may be well calibrated. These observations may depend on variability in the domain of reading, however. That is, the expert knows that he or she is competent in the domain of ex-

This research was funded by Office of Naval Research Contract No. N0014-85-K-0644 and National Institute of Education Grant No. NIE-G081-0009 to the Wisconsin Center for Education Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institute of Education or the Department of Education.

Our thanks to Craig Morris, Tom Sanocki, and Naomi Swanson for assisting in execution of this research.

Requests for reprints should be sent to either author at the Department of Psychology, W. J. Brogden Psychology Building, University of Wisconsin, Madison, WI 53706.

pertise and less competent in other domains. Thus, by using base rates, the expert can accurately predict better performance in the domain of expertise than in alternative domains. Nonetheless, this ability to predict relative performance across domains does not imply that the expert is well calibrated within a domain.

In fact, a sampling of the literature indicates that relative expertise does not confer an ability to predict performance within the domain. Oskamp (1965) has reported that trained clinical psychologists are greatly overconfident in their predictions derived from reading case studies. Similarly, Hock (1985) found that students in a master's in business administration program were overconfident in their predictions of their future success in developing employment opportunities. Bradley (1981) had undergraduates rank their knowledge in 12 domains. He then administered a short test on content from each domain and had subjects rate confidence in each answer. Performance on the test was positively related to the knowledge rankings. However, confidence in incorrect answers also increased with the knowledge ranking. The "experts" were less likely (or willing) to admit ignorance.

We recruited subjects who had a minimum of two college-level physics courses or two college-level music courses (excluding performance courses, such as marching band). Within each of these groups, subjects had a wide range of formal course work and nonacademic experience. We chose these two domains because they have little overlap. Also, Birkmire (1982) found that music students were more sensitive to structurally important components of the text when reading in the domain of music than when reading in the domain of physics. Physics students showed the converse effect.

Our stimulus materials were prepared by two advanced graduate students: a graduate student in physics composed 16 expositions on various topics in physics; a graduate student in music theory composed 16 expositions on various topics in music. The texts were written to be representative of introductory exposition in each domain. Each of the subjects read all of these texts, 8 physics texts and 8 music texts on each of 2 days. At the end of each day's session, the subject rated confidence in his/her ability to correctly answer inferences for each text and was given the inference verification test. (Glenberg & Epstein, 1985, demonstrated that delaying the confidence assessment until the end of a session does not change calibration.)

The expertise hypothesis predicts that physics students will be better calibrated for the physics texts than for the music texts, and that music students will show the opposite pattern. On the other hand, expertise may only confer the ability to predict better performance (not better calibration) in the domain of expertise than in an alternative domain. In the latter case, (1) experts will be poorly calibrated within both domains, but (2) calibration computed across domains will be greater than zero.

The experiment was also designed to examine a number of other issues. Glenberg and Epstein (1985) found that although the average measure of calibration was not significantly different from zero, there was large variation in the point-biserial correlations. Having subjects read texts on 2 days allowed us to determine whether this variability is due to random error or to stable individual differences.

In addition to obtaining information from subjects regarding their experiences in the domains of physics and music, we assessed each subject on the dualism scale (Ryan, 1984). A dualist, who has relatively immature epistemological standards, believes that truth is absolute in most if not all domains. A relativist believes that truth is determined by the context and that propositions are true or false within a particular frame of reference. Ryan demonstrated that relativists engage in more sophisticated comprehension monitoring than do dualists. We wished to determine whether stable individual differences in calibration are associated with these different epistemological styles.

The experiment was also designed to test the generality of two other findings reported by Glenberg and Epstein (1985). In Glenberg and Epstein's third experiment, subjects provided three responses after answering the inference question for each text. First, the subject was asked to rate confidence in the correctness of the answer to the inference question. The correlation of this confidence rating and performance on the test is called *calibration of performance*. In contrast to initial calibration, calibration of performance was significantly greater than zero. This finding is consonant with Lichtenstein, Fischhoff, and Phillips's (1982) results that accuracy of postdictions are significantly better than chance (although generally exhibiting overconfidence).

After rating confidence in performance, subjects in Glenberg and Epstein's (1985) third experiment provided another assessment of confidence in their ability to judge inferences on an upcoming test. Then a second inference test was given. The correlation between this second prediction and performance on the second test is called recalibration. In Glenberg and Epstein's third experiment, recalibration was significantly greater than zero. Glenberg and Epstein proposed that the experience (e.g., ease of retrieval of relevant propositions, amount of time required to check the inference) gained from answering the first inference question provided valid cues to the degree of comprehension, and that these cues could be used to predict future performance. A similar hypothesis has been offered to explain the relationship between accuracy and confidence in eyewitness identification. Kassin (1985) found that subjects in the eyewitness identification task are generally poorly calibrated. Having subjects attend to the experience of making a judgment results in significant improvements in calibration.

The current experiment includes the measurements needed to compute both calibration of performance and recalibration. Either of these measures may be related to expertise in a domain of knowledge.

In preview, the results met some of our expectations but not others. Generally, calibration within a domain was

low, especially for experts. On the other hand, calibration of performance (judging correctness of already answered questions) was quite high, indicating the reliability of the measures. Also, calibration across domains was high. Internal analysis suggests that this last finding reflects the subjects' self-classification as expert or not (within different domains), rather than the subjects' ability to accurately assess knowledge gained from particular texts.

Method

Subjects

Seventy subjects were recruited from the University of Wisconsin-Madison community. Several recruitment procedures were used, including posters advertising the experiment, mailings to students meeting the minimum course work requirements, and solicitation in upper-level classes. The minimum course work requirement was completion of two university-level courses in either physics or music theory. Upon completing the experiment, subjects completed a questionnaire requiring a listing of the university-level music and physics courses completed, as well as a listing of other experiences in music (e.g., lessons in playing an instrument) or in physics (working as a laboratory assistant). These experiences were coded using a scale of 0 (*no experience*) to 3 (*experience at a professional level*, such as giving music lessons).

Some subjects were eliminated from the data analyses. Six were eliminated from all analyses for failing to provide complete background information. Three were eliminated from all analyses because calibration measures could not be computed, either because of perfect performance on all test questions or because the subject always used the maximum confidence rating. Finally, 11 subjects were eliminated from at least one analysis when calibration for a particular analysis could not be computed.

Table 1 provides descriptive statistics for the 50 subjects included in all analyses. The data on both music and physics courses indicate that a large range of experience is represented in the analyses. Of the 14 subjects with complete background data who were eliminated, none had background data outside the ranges given in Table 1. For those 14 subjects, the mean number of music courses was 3.36 and the mean number of physics courses was 3.14. These means are a little higher than the means in Table 1, indicating that the subjects who were eliminated were somewhat more expert than those remaining in the analyses. However, the differences in the means are small, within .25 standard deviations.

Since there were subjects who had relevant experience in both music and physics, we did not attempt to classify subjects into mutually exclusive categories. Instead, background knowledge was coded using four variables: number of music courses, music experience, number of physics courses, and physics experience. These four variables were then entered, as a set, into a hierarchical multiple regression analysis to determine the effect of background knowledge on calibration.

The questionnaire also contained a 7-item scale for measuring dualism (Ryan, 1984). Subjects rated the relative frequency (1 = *rarely*, 5 = *almost always*) of experiencing such thoughts as "If professors would stick more to the facts and do less theorizing one could get more out of college." The higher the average rating, the greater the tendency toward dualism. Data from this scale are also given in Table 1.

Subjects were paid $8 for participating in the experiment.

Table 1
Subject Characteristics

Variable	Mean	*SD*	Smallest	Largest
Dualism	2.59	0.80	1.14	4.14
Music Courses	2.76	3.77	0.00	15.00
Music Experience	1.34	0.96	0.00	3.00
Physics Courses	2.56	2.38	0.00	11.00
Physics Experience	0.26	0.49	0.00	2.00

Materials

Each text was one paragraph long and was written to illustrate or explicate a central principle that was stated explicitly in the text. An example is presented in the Appendix with the central principle italicized. The principle was not emphasized for the subjects. Two pairs of inference questions were written for each text. One member of each pair was a true inference, and the other member of each pair was a false inference. A particular inference verification item consisted of one member of one pair that the subject judged as true or false. Accurate performance on the inference tests required knowledge of the central principle. Assignment of an inference pair to the first or second inference test was counterbalanced. Also, choice of the true or false inference for each pair was counterbalanced, and the correct answer on the second test was independent of the correct answer on the first test. Examples of the inference tests are provided in the Appendix.

The texts were arranged in two booklets with 16 texts in each. One booklet was used for the first session, and one booklet was used for the second. Within each booklet, there were 8 music texts alternating with 8 physics texts. The order of the texts was counterbalanced across subjects.

Following the texts in each booklet were 16 sets of five probes. Each set corresponded to one of the texts, and the sets were in the same order as the texts. The confidence probe (Probe 1) gave the title of the text and required the subject to indicate confidence in his/her ability to judge the correctness of an inference regarding a reference to the central principle (see the Appendix for examples). The subject responded by circling a confidence rating of 1 (*very low*) to 6 (*very high*).

The inference test (Probe 2) was on the following page (headed by the title of the relevant text). Each subject judged the correctness of the inference by circling a T (*true*) or F (*false*). The confidence in performance scale (Probe 3) was on the same page. The subject was asked to rate his/her confidence that he/she had answered the inference test correctly (using a number from 1 to 6). The recalibration confidence scale (Probe 4) was also on this page. The subject indicated confidence in his/her ability to answer another inference regarding the central principle. Once again, confidence was indicated by circling a number from 1 to 6.

The following page presented the second inference test (Probe 5). This page was also headed by the title of the text. Again, each subject responded by circling T or F.

Procedure

Subjects were tested in small groups. The instructions explained that the aim of the experiment was to investigate how students assess comprehension. They were told that they could read the passages at their own pace, and rereading of a passage was allowed. However, once any page was turned, it could not be turned back. Further instruction regarding how to answer the five probes was also provided.

On Day 1, the experiment was adjourned after subjects had read and completed the 16 sets of probes. The second session was scheduled for 1 to 7 days later. At the end of the second session, the subjects completed two questionnaires. For the first, subjects were asked to rate familiarity with the topics of each of the 32 texts on a scale from 1 to 6. Subjects were provided with copies of the texts while producing the ratings. The second questionnaire was the survey of domain-specific experiences and dualism.

RESULTS

Due to the continuous nature of the background knowledge variables, the basic strategy of data analysis was to use hierarchical multiple regression techniques to perform an analysis of variance (Cohen & Cohen, 1975). Two groups of analyses were performed. In the initial analyses, the between-subjects variables were dualism (entered into the regression first) and the four background knowledge variables (entered as a set with four degrees of freedom). The protected-*t* procedure was used; the significance of individual components of the background knowledge set was examined only when the omnibus *F* was significant. The within-subjects variables were type of text (music or physics) and the interaction of type of text and background knowledge. The protected-*t* procedure was also used to examine components of this interaction. The interaction of dualism and type of text was not examined. The *MS*e terms were computed by dividing the proportion of (between-subject or within-subject) variance not accounted for by any of the independent variables by the degrees of freedom.

The second set of analyses was motivated by two concerns. First, the dualism variable accounted for little variance and, thus, tended to waste degrees of freedom. Second, there were significant positive correlations between music experience and music courses variables (.62) and between physics experience and physics courses variables (.47). These correlations can distort the significance levels of the individual variables when they are entered as a set (the problem of collinearity, Cohen & Cohen, 1975). For these reasons, the second set of analyses omitted the dualism, music experience, and physics experience variables. Fortunately, the second set of analyses produced a very similar pattern of significant results as did the first set of analyses. Because the second analyses are simpler, they will be the main focus of the results discussed here. Reference to the first analyses will only be made when there is a significant discrepancy between the two.

The measurement of calibration requires variability in both the use of the confidence scale and in performance on the inference test. Because some subjects used the same confidence level for all texts or answered all of the inference questions correctly, they were excluded from some of the analyses. Consequently, the number of subjects contributing to each analysis differed. This number is indicated at the beginning of each of the sections dealing with separate analyses.

The Type 1 error rate was set at .05 for all analyses.

Initial Calibration and Its Components

Confidence (Probe 1), $n=61$. The mean confidence on the music texts was 4.69 (SD = .99), and the mean confidence on the physics texts was 4.73 (SD = .94). These means were not significantly different. There was one significant effect in the analysis of variance: type of text interacted with background knowledge [$F(4,116)$ = 79.34, MSe = .0024]. Both of the background knowledge variables, number of music courses and number of physics courses, were significant contributors to this interaction.

Table 2
Regression Coefficients for Initial Calibration and Its Components

Type of Text	Independent Variable: *Y* Intercept	Music Courses	Physics Courses
	Confidence		
Music Text	4.7471	0.1003[a]	−0.1300[b]
Physics Text	4.5301	−0.0789[a]	0.1601[b]
	Proportion Correct		
Music Text	0.6453[c]	0.0121[d]	0.0159*
Physics Text	0.7275[c]	−0.0022[d]	0.0275*
	Goodman-Kruskal *G*		
Music Text	0.1034	−0.0251	0.0120[e]
Physics Text	0.3740	−0.0213	−0.1165[e]

Note—*The coefficients of variables having significant main effects (significantly related to the dependent variable averaged over text type). Coefficients with the same letter are significantly different from one another and indicate that a significant interaction exists between the independent variable and text type.

The regression coefficients are given in Table 2. A significant effect indicates a significant increment in proportion of variance accounted for by addition of the independent variable to the multiple regression equation. It is closely analogous to a significant effect in an analysis of variance. Furthermore, if one is willing to treat the variables as interval measures, then these coefficients indicate the average change in the dependent variable (in this case, confidence) for each unit change in the independent variable.

The coefficients in Table 2 indicate a reasonable pattern of relationships between the independent variables and confidence. Confidence in music texts increases with the number of music courses, and the increase for music texts is significantly greater than the increase for the physics texts. Also, confidence in physics texts increases with number of physics courses, and that increase is significantly greater for the physics texts than for the music texts.

These results provide manipulation checks on the construction and classification of the texts and on the validity of the background knowledge variables. That is, the interaction between text type and confidence is what would be expected if our subjects did indeed differ in expertise in the two fields, and the texts tapped that difference.

Proportion correct on the first inference test (Probe 2), $n=61$. Mean proportion correct was .72 (SD = .12) on the music texts and .79 (SD = .12) on the physics texts, a significant difference [$F(4,116)$ = 38.39, MSe = .0021]. The set of background knowledge variables also accounted for a significant part of the variance [$F(2,58)$ = 8.48, MSe = .0133]. Only the physics courses variable was significant by the protected-*t* procedure. Each additional physics course was associated with a .0217 increase in proportion correct (averaged over both types of text).

In the first analysis of proportion correct, a significant main effect was found for dualism [$F(1,55) = 4.54$, $MSe = .0129$]. Each unit increment on the dualism scale was associated with a .0268 reduction in proportion correct.

There was also a significant interaction between type of text and background knowledge [$F(2,116) = 19.42$, $MSe = .0021$]. The regression coefficients for this interaction are given in Table 2. The major component carrying the interaction was number of music courses. Proportion correct on the music texts increased with increases in music courses, whereas proportion correct on the physics texts was essentially unrelated to music courses. The opposite pattern was found for the physics courses variable (although not significant): Proportion correct on the physics texts increased more with related experience than did proportion correct on the music texts. The failure to reach significance may in part reflect the problem of collinearity. The two variables are significantly, although negatively, correlated (−.44).

Calibration of comprehension, $n=50$. Calibration is a measure of the degree of association between confidence and performance on the inference test. One such measure is the point-biserial correlation. Unfortunately, this measure has a number of undesirable properties, including the fact that the maximum value depends on the proportion correct. Nelson (1984) suggests that the Goodman-Kruskal gamma (G) is the most appropriate index of association for measuring metacognitive performance under the conditions instantiated in this experiment. Gamma ranges from −1 to 1, with 0 indicating no relationship. It has a direct interpretation in terms of the difference between two probabilities. Consider all pairs of texts that differ on both confidence and performance on the inference test for a given subject. Gamma is the difference between the probability that the text with the greater confidence has the better performance and the probability that the text with the greater confidence has the lower performance.

For each subject, G was computed separately for the music texts and for the physics texts. In each domain, G was computed from each subject's confidence ratings and performance. The means were .06 ($SD = .53$) for the music texts and .02 ($SD = .62$) for the physics texts. Neither of these means was significantly different from zero, and they were not significantly different from one another.[1] Although none of the main effects was significant, there was a significant interaction between type of text and background knowledge [$F(2,94) = 7.99$, $MSe = .0044$]. The regression coefficients for this interaction are given in Table 2. The significant component of the interaction was the interaction of text type and number of physics courses. An increase in number of physics courses tended to decrease G for the physics texts, but had essentially no relationship to G for the music texts.

The finding of no overall calibration of comprehension replicates the previous results of Glenberg and Epstein (1985). The new information provided by the present experiment concerns the relationship between level of knowledge in a domain and calibration in that domain. Under these experimental conditions, that relationship is negative. Note that for the physics texts, subjects with no physics courses and the average number of music courses (2.76) are predicted by the regression equation to be fairly well calibrated ($G=.3152$). However, the predicted G drops to .0170 for subjects with the average number of both music and physics courses. This new result is discussed further in the Discussion.

Calibration of Performance

Confidence in performance (Probe 3), $n=61$. After answering an inference question, a subject rated confidence in his/her answer to the inference question. The mean confidence ratings were 4.76 ($SD = .73$) and 4.99 ($SD = .67$) for the music and physics texts, respectively. These means were significantly different [$F(1,116) = 12.22$, $MSe = .0021$]. There was also a significant interaction between type of text and background knowledge [$F(2,116) = 59.59$, $MSe = .0021$]. Each of the background knowledge variables contributed to this interaction ($ts > 3.65$).

The regression coefficients are given in Table 3. Note that the pattern of the coefficients differs for confidence (Probe 1, Table 2) and confidence in performance (Probe 3, Table 3). That is, for both variables, the difference between the coefficients for music texts and physics texts is smaller in Table 3 than in Table 2. We will use this difference to argue (in the Discussion) that subjects used different strategies to produce the two confidence ratings.

Calibration of performance (Probes 2 and 3), $n=55$. Is there a significant relationship (G) between confidence in performance and actual performance? In short, the answer is yes. The average performance G for the music texts was .42 ($SD = .43$), and the average performance G for the physics texts was .36 ($SD = .55$). Both of these Gs are significantly greater than zero, and they are sizable on an absolute scale. Remember that G is a difference in probabilities: An average G of .39 means that for texts that differ in confidence and correctness on the inference test, the probability that the text with the greater confidence is correct is .39 greater than the probability that the text with the lower confidence is correct.

Performance G was unrelated to number of music courses and unrelated to number of physics courses; also, the background knowledge variables did not interact with

Table 3
Regression Coefficients for Performance Confidence and Calibration

	Independent Variable		
Dependent Variable	*Y* Intercept	Music Courses	Physics Courses
Music Text Confidence	4.7179[a]	0.0775[b]	−0.0671[c]
Physics Text Confidence	4.8523[a]	−0.0377[b]	0.0910[c]
Average G	0.4517	−0.0081	−0.0154

Note—Coefficients with the same letters are significantly different from one another and indicate that a significant interaction exists between the independent variable and text type.

type of text. Thus, to the extent that the null hypothesis is supported, calibration of performance is unrelated to expertise.

The significant performance *G* is important in two respects. First, it replicates a previous finding of Glenberg and Epstein (1985), and creates a bridge between our work on calibration of comprehension and other work on calibration of probabilities. The ability to accurately postdict performance has been a stable feature of the calibration literature (Lichtenstein et al., 1982). Second, the significant performance *G* helps to rule out some uninteresting interpretations of the nonsignificant calibration of comprehension *G*. In particular, given that performance *G* is significant, it is less likely that the nonsignificant calibration of comprehension *G* reflects low statistical power or any hidden constraints in our procedures.

Recalibration and Its Components

Recalibration confidence (Probe 4), $n=61$. After assessing confidence in performance, subjects were asked for confidence in their ability to answer a second inference test related to the same principle. Recalibration confidence is markedly similar to calibration confidence (Probe 1). The recalibration confidence means were 4.67 (*SD* = .87) and 4.72 (*SD* = .88) for the music and physics texts, respectively. The only significant effect was the interaction of text type and background knowledge [$F(4,116) = 77.14$, *MS*e = .0022]. The regression coefficients are given in Table 4. Note that for both variables, the difference between the coefficients for the music and physics texts is almost as great for recalibration confidence as for calibration confidence (Table 2).

Recalibration proportion correct (Probe 5), $n=61$. Performance on the second inference test was similar to performance on the first. The mean proportions correct were .73 (*SD* = .13) and .79 (*SD* = .12) for the music and physics texts, respectively. The difference was significant [$F(1,116) = 21.48$, *MS*e = .0030].

Table 4
Regression Coefficients for Recalibration and Its Components

	Independent Variable		
Type of Text	*Y* Intercept	Music Courses	Physics Courses
	Confidence		
Music Text	4.6512	0.0944[a]	−0.0961[b]
Physics Text	4.5287	−0.0667[a]	0.1421[b]
	Proportion Correct		
Music Text	0.7048[c]	0.0060	0.0012[d]
Physics Text	0.7301[c]	0.0000	0.0224[d]
	Goodman-Kruskal *G*		
Music Text	−0.0596	0.0309*	0.0098[e]
Physics Text	0.1768	0.0277*	−0.0918[e]

Note—*The coefficients of variables having significant main effects (significantly related to the dependent variable averaged over text type). Coefficients with the same letter are significantly different from one another and indicate that a significant interaction exists between the independent variable and text type.

There was also a significant interaction between type of text and background knowledge [$F(2,116) = 10.61$, *MS*e = .0030]. The regression coefficients are listed in Table 4. The only significant component in the interaction involves the number of physics courses variable. Increments in number of physics courses are associated with increments in proportion correct for the physics texts, but not for the music texts. (This effect was not significant in the first analysis using four variables to code background knowledge.)

As in the analysis of the first inference test, there was a main effect for dualism [$F(1,55) = 8.15$, *MS*e = .0135] in the first set of analyses. On the average, a unit increase in the dualism variable was associated with a decrease of .0365 in proportion correct.

Recalibration *G*, $n=54$. In each domain, a recalibration *G* was computed from each subject's confidence rating on Probe 4 and performance on the second inference test (Probe 5). Mean calibration *G*s were .06 and .02 for the music and physics texts, respectively. Neither was significantly different from zero. Background knowledge did account for a significant proportion of the variance in recalibration *G* [$F(2,51) = 4.49$, *MS*e = .0167]. Number of music courses was the variable that contributed most significantly.

There was also a significant interaction between type of text and background knowledge [$F(2,102) = 6.12$, *MS*e = .0032] that was carried by the physics courses variable. The regression coefficients for this interaction are given in Table 4. As with initial calibration, increments in physics courses had a greater detrimental effect on recalibration for the physics texts than for the music texts.

The recalibration data do not replicate the effect reported by Glenberg and Epstein (1985), who found that recalibration was significantly greater than initial calibration (based on Probes 1 and 2). In the present experiment, overall recalibration is not different from zero, and any effect of expertise is to decrease recalibration, much as it decreases initial calibration. This failure to replicate is addressed in the discussion.

Stability of Calibration Over Days, $n=61$

Two new calibration *G*s were computed for each subject, one for Day 1 and one for Day 2 of the experiment. Each of these *G*s was based on Probes 1 (initial confidence) and 2 (initial inference evaluation) for 16 texts, 8 music texts and 8 physics texts. (All previously reported *G*s were computed separately for different types of texts).

The across-text-type *G*s were .18 (*SD* = .54) and .30 (*SD* = .45) for Day 1 and Day 2, respectively. Both of these *G*s are significantly greater than zero (*t*s = 2.60 and 5.21, respectively).

The correlation between across-text-type *G* for Day 1 and across-text-type *G* for Day 2 was only −.03. This may be compared with the correlation between confidence (Probe 1) on Day 1 and that on Day 2 (.84), and the correlation between proportion correct on each of the two days (.37). This failure to find stable individual differ-

ences suggests that the search for variables (e.g., dualism) that would correlate with calibration is futile.

These data present somewhat of a mystery. Why should *G* computed by collapsing across type of text be significantly greater than zero, when calibration (based on the same number of texts) computed within a type of text is essentially zero? One rather uninteresting explanation is that *G* based on a single type of text suffers from a restricted range; combining across text types pools texts that have a greater range on both the confidence scale and proportion correct, resulting in a larger *G*.

Two arguments can be made against this explanation. First, *G*, unlike the product-moment correlations, requires only ordinal data. In fact, theoretically, the value of the statistic is completely unaffected by the range of confidence scores, as long as there is some variability so that the statistic can be computed.

Second, performance *G*s were significantly greater than zero. These performance *G*s use the same proportion correct data as the calibration *G*s that are not significantly different from zero. Clearly, the poor calibration *G*s cannot be attributed to restricted range of performance.

A second explanation for the significant across-text-type *G*s is provided by the following hypothesis. We suppose that subjects can accurately classify themselves as relatively more expert in music or in physics. We also suppose that self-classified music students believe that they will do better on music texts than on physics texts, and that self-classified physics students believe the opposite. In fact, these beliefs are consistent with the results of our analyses of proportion correct. Finally, we suppose that confidence is based on these general beliefs, rather than on experience with the particular texts. Because performance is better in texts in the domain consonant with the self-classification than in the other domain, the self-classification is indeed predictive of performance, so that across-text-type *G* is greater than zero. According to this hypothesis, calibration across domains simply reflects the expert's use of base rates to accurately predict differences in performance across domains.

There is strong evidence consistent with the self-classification hypothesis. According to the hypothesis, subjects use their background experience with music or physics (rather than textual experience) to generate a confidence assessment for each text. This experience is public data, at least to the extent that it is revealed on the questionnaire filled out at the end of the experiment (see Method and Table 1). If the hypothesis is correct, we should be able to use these public data to generate confidence ratings that predict performance as well as the confidence ratings actually given by the subjects.

The test of this prediction required several steps. (A total of 43 subjects contributed to all steps.) First, a calibration *G* was computed for each subject using all 32 texts (to provide a maximally sensitive test). The average *G* was .20 (SD = .35), which is significantly greater than zero (t = 3.75). Next, we computed for each subject a single simulated confidence rating for all music texts and a single simulated confidence rating for all physics texts. These simulated confidence ratings were computed by entering a subject's background data (number of physics courses and number of music courses) into the regression equations specified by the coefficients for confidence given in Table 2. Finally, we computed a simulated *G* for each subject using the simulated confidence ratings in place of the actual confidence ratings.

The mean simulated *G* was .22 (SD = .44). This *G* was significantly greater than zero (t = 3.28). The mean simulated *G* and the mean of the actual *G*s (based on 32 texts) were not significantly different. Importantly, the correlation between the simulated *G*s based on public data and the *G*s based on the subjects' own 32 confidence ratings was .57.

An implication of the self-classification hypothesis is that subjects generated confidence assessments without using any sort of privileged access (Lovelace, 1984) to their own knowledge regarding the specific texts. Indeed, the hypothesis implies that subjects are not assessing comprehension of the texts at all; instead they are simply recording a belief based on their general experience. Thus, the significant across-text-type *G* should not be taken as evidence of accurate self-assessments comprehension. As just demonstrated, the confidence scores generated by the regression equation, which obviously has no privileged access to the subject's degree of comprehension of individual texts, can predict performance as well as the subject's own confidence ratings.

A similar explanation can be applied to the significant correlation between average confidence and average performance. On Day 1 the correlation was .51, and on Day 2 the correlation was .37. These correlations do not imply that subjects are calibrated. Some subjects know that they generally do well on tests and, hence, have high confidence; other subjects know that they generally do poorly on tests and, hence, have low confidence. To the extent that past experience predicts future performance, there is a correlation between average confidence and performance. However, neither the subjects who generally do well nor those who generally do poorly can accurately assess comprehension and predict which inference tests will be answered correctly: When calibration must be based on actual assessments of comprehension (i.e., within a text type), calibration is close to zero.

DISCUSSION

This experiment was designed to answer four questions. The first question was whether calibration of comprehension for texts in a given domain changes with expertise in that domain. The answer is yes, but perhaps in an unexpected way. The regression analyses for both calibration and recalibration indicate that *G* decreases with experience in a domain (and significantly so for physics).

The second question was whether there are stable in-

dividual differences in calibration of comprehension. Here the answer is no. Even the significant across-text-type *G* was not stable across days.

The third question was whether accurate calibration of performance would be found. For this question the answer is yes. Calibration of performance was not only statistically significant, it was quite large (.42 for the music texts and .36 for the physics texts; recall that *G* is the difference between two probabilities). Apparently, subjects can fairly accurately judge the quality of their performance on an inference verification test.

The fourth question concerned recalibration. Previous results indicated that subjects could take advantage of experience gained while answering an inference test to predict performance on future tests over the same material. The subjects participating in this experiment did not exhibit this ability.

Self-Classification Hypothesis

The pattern of the results is consistent with the self-classification hypothesis. The hypothesis is that subjects classified themselves as relatively expert in music or physics and used the belief that expertise in a domain is correlated with comprehension of texts in that domain to generate confidence ratings. That is, self-classification rather than assessment of text comprehension controlled the confidence ratings.

The strongest evidence consistent with the hypothesis is from the analysis of the simulated *G*s. The mean simulated *G* was not significantly different from the mean *G* produced by the subjects, and the correlation between the simulated *G*s and the actual across-text-type *G*s was substantial.

The self-classification hypothesis provides a simple explanation for the poor calibration within a text type. According to the hypothesis, subjects are not actually assessing knowledge gained from a particular text; instead they are responding on the basis of beliefs about their abilities within a given domain. These beliefs are not sufficiently fine-grained (differentiated) to accurately predict performance within a domain.

Variability of confidence ratings within a domain may be based on judged familiarity with a topic. In fact, the average Pearson correlation between familiarity ratings (obtained at the end of the second session) and confidence was .63 (SD = .17). When these familiarity ratings (one for each text) are used to compute a *G*, the average familiarity G = .23 (SD = .29), which is not significantly different from the average simulated *G* based on a single confidence rating for each type of text. Thus, although the familiarity ratings account for variability in the confidence ratings, they do not contain any useful information for predicting performance over and above that provided by the self-classifications.

The self-classification hypothesis is also at least partially consistent with the negative relationship between expertise and calibration (within a domain). Most likely, only subjects who regard themselves as having some expertise will apply the self-classification strategy. Other subjects may actually carry out some form of evaluation of comprehension that predicts performance on the inference test. (Based on the regression equations, subjects with an average number of music courses, but no physics courses, were calibrated.) Thus, increasing expertise is associated with application of a less successful strategy for predicting performance within a domain.

The self-classification strategy was probably also applied when subjects were asked to re-assess confidence (Probe 4) in future performance. The pattern of regression coefficients relating background knowledge to initial confidence (Probe 1) was similar to the pattern relating background knowledge to re-assessed confidence (Probe 4, compare Tables 2 and 4). Apparently subjects were using the same information (self-classifications) to make both ratings.

On the other hand, it appears that confidence in performance (Probe 3) was not determined by self-classification. First, these confidence ratings were significantly correlated with actual performance (performance *G* greater than zero) within a domain of knowledge, which is not possible by application of the self-classification strategy alone. Second, the pattern of regression coefficients relating background knowledge to confidence in performance is quite different from the pattern relating background knowledge to initial confidence (compare coefficients in Table 3 to those in Table 2).

When is the Self-Classification Strategy Applied?

We have stressed the contribution that self-classification may make to the computation of confidence. But we do not intend to imply that the metacognitive rule expressing the relationship between self-classification and likelihood of successful performance is the only rule for computing confidence. Other rules based on familiarity and ease or completeness of access to the relevant text may also be engaged. This suggestion is consistent with the significant correlation between familiarity ratings and confidence ratings.

Given that there is a repertoire of metacognitive rules for computing confidence, when is the self-classification strategy applied? One consideration may be the task setting. Various features of the setting of the current experiment probably encouraged use of the strategy. Subjects knew that they were selected on the basis of their experience in music and physics courses. In addition, the texts were clearly in one domain or the other, and the contrast was heightened by the presentation order, which alternated texts from the two domains. Probably, the strategy is encouraged whenever the domain of the text clearly matches the subject's own beliefs about domains of expertise.

In addition to the task setting, other factors affecting availability of rules in memory may be involved in determining the subject's choice from the repertoire of

metacognitive rules. Also, it seems likely that the process of selection is dynamic, reflecting the effects of several variables operating concurrently to assign prominence to different metacognitive rules. The dynamic character of the process helps us to formulate a coherent account of the principal findings of this study.

We have argued that the initial confidence rating was computed by application of the self-classification strategy, the rule made most available by the task setting. Why then, was the self-classification strategy not applied by subjects when rating confidence in performance? After answering the first inference test (Probe 2), subjects could base their confidence ratings on either the self-classification strategy or the specific experience gained from answering the inference (such as ease of retrieving relevant propositions from memory). We propose that most subjects chose to use specific experience for the following reasons: (1) Because the subjects had just evaluated the inference (Probe 2), the experience was probably highly available while they made the confidence in performance rating (Probe 3). (2) Some of the specific experiences were probably easily recognized as diagnostic. For example, failure to retrieve any information relevant to evaluating the inference is easily recognized as a useful predictor of chance performance. (3) The experience was specific to the particular judgment being made, whereas the self-classification strategy is more general.

On the other hand, it appears that the self-classification strategy was applied again in generating predictions about future performance on the recalibration confidence rating (Probe 4, see discussion of recalibration). Why do subjects revert to using the self-classification strategy for Probe 4, after rejecting it for Probe 3? In answering Probe 4, subjects also have a choice of metacognitive rules. We suspect that the self-classification strategy is chosen because of a difference in the diagnostic value attributed by the subject to the experience gained from answering the initial inference. Experience answering the first inference is believed to be diagnostic for judging performance on the first inference. The experience is believed to have less diagnostic value for predicting future performance. Given the belief that the diagnostic value of the experience is low, and the ready availability of a strategy with high face validity, subjects chose the self-classification strategy.

Subject's use of the self-classification strategy when answering Probe 4 helps to explain why significant recalibration was not found in this experiment, but was found in previous experiments (Glenberg & Epstein, 1985). As discussed before, the self-classification strategy cannot produce calibration within a domain, obviating any possibility of significant recalibration. Glenberg and Epstein (1985) sampled texts from a variety of domains, reducing availability and use of the self-classification strategy. Thus, in that previous research, when subjects re-assessed confidence after the initial inference test, it is likely that the subjects were forced to use a metacognitive rule with greater predictive validity than the self-classification strategy.

In summary, it appears that the self-classification strategy will be used (and will be effective) under the following conditions. First, the structure of the calibration task (e.g., using texts from two domains) suggests the strategy by highlighting the relationship between a reader's domain of knowledge and the domain of the text. Second, the reader does not have available information that is believed to be more specific or more diagnostic than self-classification. Whether or not application of the strategy produces calibration depends at least in part on the structure of the task. Application of the strategy across domains of expertise is almost guaranteed to produce high calibration. Unfortunately, the self-classification strategy alone cannot produce calibration within a domain of expertise.

REFERENCES

BIRKMIRE, D. P. (1982). *Effect of the interaction of text structure, background knowledge, and purpose on attention to text* (Technical Memorandum 6-82). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

BRADLEY, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, **17**, 82-84.

COHEN, J., & COHEN, P. (1975). *Applied multiple regression/correlation analyses for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

GLENBERG, A. M., & EPSTEIN, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **11**, 702-718.

GLENBERG, A. M., & EPSTEIN, W. (in press). *Enhancing calibration. Journal of Experimental Psychology: General.*

HOCK, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **11**, 719-731.

KASSIN, S. M. (1985). Eyewitness identification: Retrospective self-awareness and the accuracy-confidence correlation. *Journal of Personality & Social Psychology*, **49**, 878-893.

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 & Cognition*, **10**, 756-766.

MAKI, R., & BERRY, S. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **10**, 663-679.

NELSON, T. O. (1984). A comparison of current measures of the accuracy of feeling of knowing predictions. *Psychological Bulletin*, **95**, 109-133.

OSKAMP, S. (1965). Overconfidence in case study judgments. *Journal of Consulting & Clinical Psychology*, **29**, 261-265.

RYAN, M. P. (1984). Monitoring text comprehension: Individual differences in epistemological standards. *Journal of Educational Psychology*, **76**, 248-258.

NOTE

1. We also computed a G for each subject ($n = 54$) using the subject's confidence ratings and the average performance on the two inference questions (Probes 2 and 5). We felt that using two questions should increase the reliability of the measure. Nonetheless, average calibrations were only .06 and −.05 for music and physics texts, respectively. These Gs were not significantly different from zero or from each other.

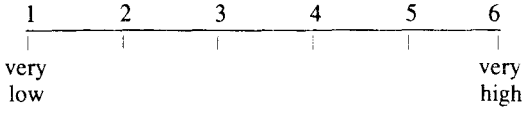
APPENDIX
Organic Unity–Text

The way in which the parts of a musical work relate to form a whole has long been an important consideration of musical aesthetics. The theory of organic unity, which directly compared the parts and whole of musical works to those of living things, became part of the evaluative process as an aesthetic norm in the early 19th century. According to the theory, musical pieces were analogous to creatures: *Each part of a successful work was essential, just as every part of the body was (supposedly) essential; no part of a good piece of music could be substituted for another, since each had a specific function in the unified whole.* Furthermore, as in an organic body, the combined functions of all the parts of a musical masterwork were believed to form a coherent unity because of specific relationships that held the parts together; thus no part of the whole could stand separately as a successful work. Certain parts of the whole were believed to carry more important functions than others, just as the heart has a more important function than the little toe. Furthermore, it was believed that great composers were great creators, who, like God, fashioned "living organisms." (Consider a statement by Karl Kahlert, music aesthetician, writing in 1848: "What is musical form but the natural body that music must assume in order to establish itself as a living organism?") Though the analogy is useful and interesting, problems with the theory of organic unity are evident. It assumed that composers were aiming at a particular kind of structural unity, which was simply not the case for most pieces written before about 1600 or after about 1910. It demonstrated an evaluative bias against longer forms, especially opera, in which the semblance of complete unity was more difficult to maintian.

Probe 1–Confidence Scale

Organic Unity

Circle a single number on the following scale to report your confidence in being able to accurately judge the correctness of an inference drawn from the reading about the relationships between parts of a composition according to the theory of organic unity.



Probe 2–Initial Inference

Organic Unity

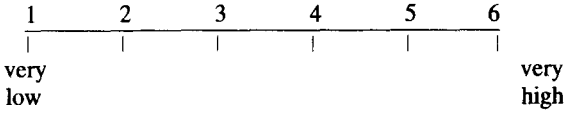
Inference: According to the theory of organic unity, it is not possible to improve some compositions by deleting specific parts.

T F

Phase 3–Confidence in Performance

Organic Unity

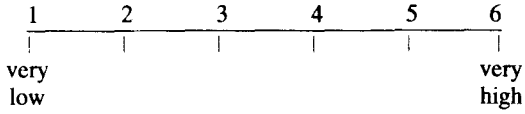
Circle a single number on the following scale to report your confidence that you have answered the inference correctly.



Probe 4–Recalibration Confidence

Organic Unity

Circle a single number on the following scale to report your confidence that you can judge the correctness of another inference drawn from the reading about the relationships between parts of a composition according to the theory of organic unity.



Probe 5–Second Inference

Organic Unity

Inference: The theory of organic unity does not explain why a single movement of a work is often complete and performable without the other movements of the compositon.

T F

(Manuscript received February 4, 1986;
revision accepted for publication June 2, 1986.)