Knowledge Monitoring Calibration: Sensitivity and Specificity as Unique Cognitive Constructs

Francis X. Smith (francis-smith@uiowa.edu)
Department of Psychology, E11 Seashore Hall,
Iowa City, IA 52242 USA

Christopher A. Was (cwas@kent.edu)
Lifespan Development and Educational Sciences, 405 White Hall
Kent, OH 44242, USA

Abstract

Knowledge monitoring is an important metacognitive process which can help students improve study habits and thereby increase academic performance. Which is more useful in predicting test performance: knowing what you know, or knowing what you do not know? Two distinct constructs of knowledge monitoring calibration, sensitivity and specificity, were used along with the more traditional Gamma to predict performance on tests in an undergraduate educational psychology course. It was found that sensitivity, a measure of correctly identifying known items, was the most useful in predicting overall test scores as well as final exam scores. Specificity, on the other hand, had no significant impact on exam performance. Results suggest that sensitivity and specificity may be more meaningful measures of knowledge monitoring calibration when it comes to predicting academic achievement, as well as being better adapted for missing values in any one cell of the data.

Keywords: knowledge monitoring, metacognition, calibration

Introduction

In the course of preparing for an examination a student must make several judgments of their knowledge. The student must decide if studying outside of lecture time is necessary to achieve the level of success desired. If studying seems appropriate the student needs to decide which materials to study and for how long. All of these decisions are based on a student's judgment of how much of the material they truly know, and will be able to recall during the exam, and how well they know it. It is, therefore, crucial that a student be able to make accurate judgments of their knowledge in order to appropriately and efficiently allocate study time and other methods of preparation.

The ability to identify what information is known and what is unknown is referred to as knowledge monitoring accuracy. It is logically reasonable to claim that for any higher-order self-regulation of learning to be effective, accurate knowledge monitoring is essential. In fact, models of self-regulated learning often include definitions such as “the setting of one’s own goals in relation to learning and ensuring that the goals set are attained” (Efklides, 2011). While it may be possible to set goals without knowledge monitoring, it would certainly be difficult to assess attainment of those goals prior to the actual evaluation without some kind of monitoring process.

A number of theories hold a similar position, arguing that effective monitoring leads to better regulation during learning (Metcalfe, 2009; Nelson & Narens, 1990). Indeed, recent evidence has supported this theoretical relationship. Nietfeld, Cao, and Osborne (2006) for example demonstrated that active practice with self-assessment throughout a semester resulted in improvements to both overall calibration (accuracy of performance predictions) as well as performance relative to another group not given the self-assessment tasks. In another recent study, it was found that effective knowledge monitoring predicted academic achievement even when the materials used to test knowledge monitoring abilities were unrelated to the material on the exams (Hartwig, Was, Isaacson, & Dunlosky, 2012). There is also some evidence that it may be possible to teach students to better monitor their knowledge (e.g., Isaacson & Was, 2010).

It seems uncontroversial to point out that these processes of monitoring one's own knowledge are only effective and beneficial if they are accurate. Research into calibration of knowledge monitoring has largely involved the use of knowledge monitoring assessments (KMA) similar to that developed by Tobias and Everson (2002). One adaptation of the format for the KMA used in prior research by Isaacson and Was (2010) is to present a series of words for the subject to identify as either known or unknown. At this point no other response is given. Importantly, the subject is not told how to process the words they are simply instructed to state if they know the meaning of the word or not. After responding to the entire list of words, subjects are then given a test to see if they can identify the meanings of each of the words out of a list of possible choices.

Effective knowledge monitoring techniques should allow an individual to successfully identify which items they know the meanings of and which items are not known. It is worth noting that, for the purposes of the KMA, the amount of items responded to correctly is not directly relevant. Rather than relying on the proportion of correct responses the results of the KMA are typically interpreted based on the proportion of items correctly identified as known or unknown. For example, if an item is identified as unknown during the initial phase and is responded to incorrectly during the testing phase this would be identified as “good” metacognitive knowledge monitoring.

The results of the KMA are generally presented in the form of a 2x2 contingency table similar to the one shown in Table 1. Cells A and D represent correctly identified items based on the responses during the initial phase and subsequent results during the test phase. Conversely, cells B and C represent misidentified items and thus inefficient or ineffective knowledge monitoring. There are a number of ways to analyze the results of the KMA regarding calibration of knowledge monitoring.

To interpret the results of the KMA, a non-parametric gamma correlation coefficient developed by Goodman and Kruskal (1954) has often been calculated (e.g., Isaacson & Was, 2010; Hartwig et al., 2012). As with any correlation coefficient, the range of values for Gamma is -1.00 to +1.00. The formula for calculating Gamma utilizes values from all four cells in both the numerator and the denominator and is written as (AD-BC) / (AD+BC). Gamma is a measure of association. Missing values can seriously impact the resulting value when calculating the Gamma coefficient.

Although Gamma is commonly used in the metacognition and knowledge monitoring literature, concerns have been raised regarding the validity and robustness of gamma as a measure of knowledge monitoring accuracy and feelings-of-knowing accuracy. In an investigation of the soundness of measures of feeling-of-knowing accuracy, Schraw (1995) originally compared gamma and the Hamann coefficient. More recently, a variety of alternatives were explored by Schraw, Kuch and Gutierrez (2012) and measures of sensitivity and specificity seem to offer a potential alternative to gamma in several important ways. In a confirmatory factor analysis, sensitivity and specificity each loaded onto independent dimensions in a two-factor model, suggesting that they may be measuring two distinct abilities that are not revealed in calculating gamma (Schraw et al., 2012). Several different models were tested including a single-factor model, a two-factor model, and a five-factor model. The two-factor model provided the best fit and in this model only sensitivity and specificity loaded strongly on the two dimensions (one each).

Sensitivity is a subject's ability to correctly identify known items among all correct responses and the formula is A / (A+C). Put differently, sensitivity is the proportion of items the subject reports knowing divided by all the items the subject responded to correctly. Specificity refers to a subject's ability to correctly identify unknown items among all incorrect responses and the formula is D / (B+D). Both are measures of diagnostic efficiency as reported in logistic regression analysis. In many ways, the comparison to logistic regression makes a great deal of sense. In logistic regression, a number of variables are used to predict a binary outcome. In knowledge monitoring, an individual may make use of a number of different internal criteria (variables) to decide if they know or do not know a piece of information (binary outcome).

In the present analysis, sensitivity and specificity are employed as predictors of academic achievement in much the same way that gamma has been used to predict achievement in similar settings (e.g., Hartwig et al., 2012). If sensitivity and specificity do represent unique constructs of metacognitive knowledge monitoring then they should be able to account for unique variance beyond that simply captured by gamma during analysis. They should also make contributions independent from one another if both constructs are important in the prediction of academic achievement. It seems intuitively plausible to argue that an individual’s ability to discriminate what they know as well as what they do not know should both make important contributions to the knowledge monitoring process. To our knowledge this is the first study to use these measures along with the KMA to study the impact of these two distinct aspects of knowledge monitoring on academic achievement.

The primary question in this study, then, is whether sensitivity and specificity will serve as better predictors of academic achievement than the more typically reported gamma. If sensitivity and specificity do serve as better predictors of academic achievement the next issue is to determine which measure is more important to the model’s performance. Rather than trying to prove that monitoring affects performance, which has been shown repeatedly in the studies cited above, the purpose in the present study was to evaluate potential alternatives to gamma.

More important from a theoretical perspective is the hypothesis that sensitivity and specificity represent unique psychological constructs. In epidemiology, sensitivity and specificity are used to evaluate a clinical test. Sensitivity represents the ability of a test to correctly identify those patients with a disease and specificity is the tests ability to correctly identify those patients without the disease. In the case of knowledge monitoring, sensitivity is one’s ability to know when information is known, and specificity is the ability to know when information is unknown. We propose that these are independent metacognitive skills and that each will predict unique variance in academic performance above and beyond that accounted for gamma, a measure of the correlation between judgments and performance.

Methods

Participants

Undergraduate students enrolled in an educational psychology course (N = 384) at a Midwestern university participated in the study in exchange for partial fulfillment of course requirements. All students were of sophomore or junior class standing. Females made up 74.5% of the sample. Data were collected between the Fall semester of 2003 and the Spring semester of 2006.

Materials

Knowledge Monitoring Assessment. As in previous research by Hartwig, Was, Isaacson, and Dunlosky (2012), the measure used to assess subjects' accuracy of knowledge

Table 1:Example 2x2 Contingency Table for KMA Results

Feeling of Knowing	**Response Accuracy**	
	Correct	**Incorrect**
Know	[A] Hits	[B] False Alarms
Don't Know	[C] Misses	[D] Correct Rejections

Note. Gamma = $\frac{AD-BC}{AD+BC}$ Sensitivity = $\frac{A}{A+C}$ Specificity = $\frac{D}{B+D}$

monitoring was adapted from Tobias and Everson (1995). The measure used in the present study involved presenting 50 vocabulary words to subjects (33 taken from the course textbook representing material from each chapter, 17 general vocabulary items) one at a time. On the first presentation subjects were to indicate whether or not they knew the meaning of the word. Importantly, there was no instruction given as to how to determine this answer. After responding to all 50 items, subjects were given a multiple choice (5 possible responses) test on these same vocabulary words. Students were required to complete this assessment within the first two weeks of the course and it was completed in an online format.

Accuracy was computed by assigning responses to cells in a 2x2 contingency table (see Table 1). Possible combinations were items identified as known and responded to correctly (hits), items identified as known but responded to incorrectly (false alarms), items identified as unknown but responded to correctly (misses), and items identified as unknown and responded to incorrectly (correct rejections). In order to evaluate the relative predictive power of different measures, gamma was calculated along with sensitivity and specificity for each individual subject.

Final Exam. For the purposes of the present study, we operationalize academic achievement as performance on a cumulative final exam at the end of the 15-week semester. The final exam was made up of 20 true/false questions as well as 80 multiple-choice items. The item were classified as three types based on Bloom's taxonomy: 40 knowledge and comprehension questions, 40 application questions, and 20 analysis and synthesis questions. Students were allowed as much time as necessary to complete the exam. Total points possible on the final exam were 100. The mean final exam score was 72.37 with a standard deviation of 12.10.

Procedure

All sections of the course in which data collection occurred were taught by the same instructor. The course materials did not vary between sections of the course. To fulfill course requirements, subjects completed the modified knowledge monitoring assessment online within the first two weeks of the semester. Students received regular feedback on performance through weekly examinations. The final exam was administered at the end of the semester and comprehensively covered material from the entire semester.

Results

Of the 384 participants, 361 completed all measures necessary to calculate a gamma score, sensitivity score, and specificity score as well as having data available for final exam performance. This represents 6% missing data. All further analysis was conducted using listwise deletion and thus did not include any data from the 6% of participants who were missing some portion of the data. Gamma [(AD-BC)/(AD+BC)], sensitivity [A/(A+C)], and specificity [D/(B+D)] were calculated for each participant (see Table 1).

To first confirm that sensitivity and specificity were predictive of academic achievement in the present sample, linear regression was used to predict final exam performance based on sensitivity and specificity scores. Results of the linear regression indicated that sensitivity and specificity accounted for a significant amount of variance in final exam scores, $R^2 = .09$, $F(2, 361) = 18.28$, $p < .001$. Whereas sensitivity was predictive of final exam scores, $B = 24.12$, $SE = 4.23$, $t = 5.67$, $p < .001$, 95% CI [15.75, 32.49],

To confirm that gamma was predictive of academic achievement, hierarchical linear regression was used to predict final exam performance based on gamma scores. The variable gamma was entered in block 1, with sensitivity and specificity entered in block 2 to examine if these two variables could account for variance beyond gamma alone. Following this, an alternative hierarchical regression model was run in which sensitivity and specificity were entered in the first block to examine how much of the variance in academic achievement was predicted by these two variables. A second block was included in which gamma was entered to examine how much additional variance would be explained. Results from these analyses are presented in Tables 2 (gamma in Block 1) and 3 (sensitivity and specificity in Block 1).

The final model for each instantiation, regardless of which variables were entered first, is the same. In each case, the full model including all three variables was able to account for 13% of the variance in final exam performance. The amount of variance accounted for is impressive considering the adapted KMA was completed online at least 13 weeks prior to the final exam and also when considering how many factors impact test performance. The full model in both instantiations was significant, $F(3,357) = 17.31$, $p < .001$. However, as shown in both Table 2 and Table 3, in the full model the only significant predictor variable was gamma.

This large change in predictor values may in part due to multicollinearity between sensitivity (VIF = 10.42, Tolerance = .11), specificity (VIF = 9.31, Tolerance = .11), and gamma (VIF = 8.99, Tolerance = .10). It also makes meaningful interpretation challenging. The seemingly appropriate interpretation in the final model is that for every one unit increase in gamma you would expect an increase in exam score of 15.8, assuming all other predictors were held constant. However, for an increase in gamma to occur there would also necessarily have to be an increase in either sensitivity or specificity or both due to the fact that the formulas for each draw from the same 2x2 contingency table. Thus, the "full model" is only used for the purposes of evaluating variance explained by the constituent predictors and not to make statements about the individual importance of predictors. To examine the variance explained by sensitivity and sensitivity or gamma alone the first block of each instantiation must be examined. When gamma was included as the only predictor of academic achievement, the model was significant as well, $F(1,359) = 25.85$, $p < .001$. In the model containing only sensitivity and specificity the overall model was significant, $F(2,358) = 23.34$, $p < .001$, although only sensitivity was significant as a predictor. Unlike the full model, the model including sensitivity and specificity did not have problems with multicollinearity and thus it seems as though sensitivity (proportion of items correct that were also identified as known) was more important in predicting exam performance than specificity (proportion of items incorrect that were also identified as unknown).

Table 2: Multiple Regression Analysis Predicting Academic Achievement From Measures of Metacognitive Knowledge Monitoring Calibration with Gamma First (N = 361)

	Block 1			Block 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Gamma	12.66	2.49	.26**	15.80	7.24	.32*
Sensitivity				3.89	11.99	.05
Specificity				-16.80	11.44	-.22
R^2		.07			.13	
F for change in R^2		25.85			12.23	

Note. *$p < .05$, **$p < .001$, $p < .001$ for *F* values.

Table 3: Alternative Multiple Regression Analysis Predicting Academic Achievement From Measures of Metacognitive Knowledge Monitoring Calibration Sensitivity and Specificity First (N = 361)

	Block 1			Block 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Sensitivity	28.24	4.40	.38***	3.89	11.99	.05
Specificity	6.21	4.44	.08	-16.80	11.44	-.22
Gamma				15.80	7.24	.32*
R^2		.12			.13	
F for change in R^2		23.34***			4.76*	

Note. *$p < .05$. **$p < .01$. ***$p < .001$.

More relevant to the current research questions, in the first instantiation of the model in which gamma was entered first, the addition of sensitivity and specificity accounted for almost twice as much variance as gamma alone. This indicates that sensitivity and specificity account for unique variance above and beyond what is being explained by gamma. On the other hand, in the second instantiation of the model (in which sensitivity and specificity were entered first) the addition of gamma only accounted for a relatively small increase in variance explained indicating that sensitivity and specificity together include most of the variance for which gamma can account.

These results suggest that sensitivity and specificity alone are more useful in predicting final exam performance. It is worth remembering, however, that even when only sensitivity and specificity were included in the model specificity was still not significant. The current results also indicate that the three measures, when analyzed together, are redundant. It seems that either gamma or sensitivity and specificity should be included but not all three simultaneously.

Interestingly, sensitivity and specificity demonstrated a significant negative correlation, $r = -.53$, $p < .001$. This pattern of negative correlations between sensitivity and specificity is often observed in meta-analyses of studies measuring diagnostic efficiency. In the confirmatory factor analysis conducted by Schraw, Kuch, and Gutierrez (2012) there was no observed correlation between sensitivity and specificity. A possible explanation offered for such a negative correlation in those instances is the use of different thresholds in different studies (Reitsma et al., 2005).

Discussion

While the scope of the present study does not allow for generalization beyond final exam performance, there seems to be genuine reason to consider reporting sensitivity and specificity in conjunction with or instead of gamma. While the three variables should not be included simultaneously in analysis, it may still be worth reporting gamma alongside sensitivity and specificity rather than simply casting it aside. It is important to recall that gamma is a different type of measure, association, than sensitivity and specificity, which are measures of diagnostic efficiency. In the current sample, specificity was never a significant predictor of final exam performance. Theory suggests, however, that being able to identify unknown items successfully should have some impact on academic performance and so it would be unwise to dismiss specificity as unimportant on the basis of this study alone. Further research may clarify this issue.

The results presented here make it clear that sensitivity and specificity not only account for almost all of the variance in exam performance explained by gamma but also explain a large portion of variance left unexplained when only reporting gamma. At least in the case of using knowledge monitoring calibration to predict academic achievement it would seem that sensitivity and specificity are the preferred measures in terms of effectiveness. These measures of diagnostic efficiency are also less problematic when it comes to missing values.

If a single cell is missing data for an individual (either A, B, C, or D) then gamma becomes significantly harder to meaningfully interpret. If either A or D is equal to 0, gamma will be either -1 or an empty set depending on if either B or C is also 0. This does not conceptually make much sense unless both A and D are equal to 0. If A or D is a non-zero number then gamma will be falsely indicating a perfect negative relationship due to the multiplication involved in calculating gamma. A similar problem exists with 0 values for B or C, with gamma shifting to 1 or an empty set if A or D is also 0. On the other hand, if A is equal to 0 then sensitivity will be equal to 0 (or an empty set if A and C are 0). Contrary to the problems with gamma, a 0 in this case actually does make conceptual sense. If a student claims they will get 0 items correct and does respond correctly to some items their sensitivity score should, and will, be 0.

Measurement and reporting aside, the most noteworthy finding in the present study is that it appears that the ability to correctly identify known items is more predictive of academic achievement than the ability to identify unknown items, as indicated by the regression models. Because most prior research has focused on general knowledge monitoring calibration, rather than on diagnostic efficiency, there may not be a readily available explanation for this effect. Intuitively it would seem that both measures should be contributing to exam performance, as both represent measures of accurate metacognitive knowledge monitoring.

In addition, the evident negative correlation between sensitivity and specificity in the present sample suggests that students are setting arbitrary thresholds at which they judge an item to be known, and that these thresholds vary from person to person. These thresholds could also be affected by individual differences in method of judging whether or not an item is known. When a student is asked if they know an item with no further instruction they may be simply using familiarity to make their judgment, or they may be trying to recall the meaning, or they may be using some alternative method. Similarly, even when using the same methods, students will have varying levels of familiarity, or success in recall, at which they will respond to the item as known as opposed to unknown.

Although the negative correlation between sensitivity and specificity is in contrast to the lack of a correlation demonstrated by Shraw, et al. (2012), we believe our findings support their conclusion that sensitivity and specificity represent two distinct cognitive mechanisms that allow individuals to make judgments about their knowledge. Also, individual differences in the accuracy of judgments made using these mechanisms may contribute to individual differences in performance.

We were surprised that sensitivity, but not specificity, was related to better academic performance. Again, we ask you to imagine a student preparing for an upcoming exam. As the student studies she must make judgments about her knowledge and understanding of the material expected to be on the exam. The judgments are used determine if a concept is well-known and no longer needs attention, or if a concept is not understood and therefore warrants further study. Put differently, are the student's judgments sensitive enough to determine when concepts are know and specific enough to know when concepts are unknown? It was our hypothesis that specificity would be more predictive of academic performance than sensitivity. This hypothesis was based on the assumption that when students are able to accurately assess when a concept is unknown, they would then exert the necessary effort to further study those items. We assumed that this would lead to better performance. Contrary to this hypothesis, sensitivity, not specificity, was predictive of academic performance. One interpretation of this finding is that students, who accurately assess what is known, can then make decisions about what no longer needs to be studied and are therefore more efficient. An alternative interpretation is that these students are simply better test takers. When taking an exam or test students with good sensitivity can effectively assess those items that are known and dedicate mental energy or cognitive load to those items that are judged as not well known. Clearly, there is a great deal of work needed to test these interpretations.

Taken together, these results suggest that efforts to improve metacognitive knowledge monitoring should focus on helping students understand how to effectively recognize if an item is truly known as opposed to seeming familiar, for example. It also seems reasonable to suggest, in the absence of further evidence, that identifying known items may be more important for academic success than identifying unknown items. Future efforts may reveal that there are situations in which specificity, rather than sensitivity, is more important in predicting outcomes. At the present time there is not strong enough evidence to warrant exclusion of specificity from analysis. If future investigations continue to demonstrate that specificity plays no significant role when using knowledge monitoring calibration to predict various outcomes then it might be worth reevaluating this position.

Conclusion

This study continues to validate the finding that knowledge monitoring ability, even when based on materials that are not directly being tested, is predictive of academic performance (Hartwig et al., 2012). However it suffers from the same problem as the previous study in that data were collected solely from educational psychology classrooms. If these findings are to be applied more broadly to different type of materials to be learned it will be necessary for future research to include a more diverse sample of classrooms. The most significant contribution of this investigation was not merely to validate the effectiveness of knowledge monitoring in predicting achievement but rather to show that there are alternative measures to the popular gamma that may be even more predictive. If these results hold up under replication then perhaps it is time to consider including sensitivity and specificity in reports of knowledge monitoring calibration.

References

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6–25.

Goodman, L. A., & Kruskal,W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association, 49*, 732–764.

Hartwig, M., Was, C., Isaacson, R., & Dunlosky, J. (2012). General knowledge monitoring as a predictor of in-class exam performance. *British Journal of Educational Psychology*, *82*, 456–468 .doi:10.1111/j.20448279. 2011.02038.x

Isaacson, R., & Was, C. A. (2010). Believing you're correct vs. knowing you're correct: A significant difference? *The Researcher*, *23*(1), 1-12.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159–163. doi: 10.1111/ j.1467 -8721.2009.01628.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159–179. doi: 10.1007/s10409-006-9595-6

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982-990. doi: doi:10.1016/j.jclinepi.2005.02.022

Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology*, *9*(4), 321-332. doi: 10.1002/ acp.2350090405

Schraw, G., Kuch, F., & Gutierrez, A. P. (2012). Measure for measure: Calibrating ten commonly used calibration scores, *Learning and Instruction, 24,* 48-57 http://dx.doi.org/10.1016/j.learninstruc.2012.08.007.

Tobias, S., & Everson, H. (2002). *Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring.* College Board Report No. 2002-3. College Board, NY.

Wickens, T. D. (2012). *Elementary Signal Detection Theory.* Oxford University Press, NY.