EDUCATIONAL SCIENCES: THEORY & PRACTICE eISSN: 2148-7561, ISSN: 2630-5984

Received: 1 July 2019
Revision received: 20 September 2019
Accepted: 25 September 2019

www.estp.com.tr

DOI 10.12738/estp.2019.4.006 ♦ **October** 2019 ♦ 19(4) ♦ 80-87

Article

Knowledge Monitoring Calibration: Individual Differences in Sensitivity and Specificity as Predictors of Academic Achievement

Francis X. Smith
University of Iowa, USA

Christopher A. Was
Kent State University, 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 Goodman-Kruskal gamma correlation to predict performance on tests in an undergraduate educational psychology course. The gamma correlation provides a measure of how good one is at judging both items one knows and items one does not. Measures of sensitivity and specificity distinguish between the two. Students in an undergraduate educational psychology course completed a 50-word knowledge monitoring assessment to measure sensitivity, specificity, and gamma. These measures were then correlated with test and final exam scores in the 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. Further research is recommended to determine in what other situations the measures of sensitivity and specificity may be useful. Findings presented in this study can also be used to help guide attempts to improve student metacognition and strategies.

Keywords

Knowledge monitoring • metacognition • calibration

Correspondence to Christopher A. Was, PhD, Department of Psychological Sciences. 313 Kent Hall, Kent State University, Kent, OH 44242. Email: cwas@kent.edu

Citation: Smith, F. X., & Was, C. A. (2019). Knowledge monitoring calibration: Individual differences in sensitivity and specificity as predictors of academic achievement. *Educational Sciences: Theory and Practice*, *19*(4), 80 -87. http://dx.doi.org/10.12738/estp.2019.4.006

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. These decisions are based on a student's judgment of how much of the material they 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 often referred to as knowledge monitoring accuracy. It is logically reasonable to claim that for any higher-order self-regulation of learning to be effective (e.g., planning, goal setting, etc.) 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, p. 6). Although it may be possible to set goals without accurate knowledge monitoring, it would certainly be difficult to assess attainment of those goals prior to the actual evaluation without a monitoring process.

Several 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. For example, Nietfeld, Cao, and Osborne (2006) 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 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., Al-Harthy, Was, & Hassan, 2015; Isaacson & Was, 2010). There is also evidence that students may not improve their knowledge monitoring with practice (e.g., Foster, Mueller, Was, Rawson, & Dunlosky, 2019).

It seems uncontroversial to point out that the processes of monitoring one's own knowledge are only effective and beneficial if they are accurate (Lichtenstein & Fischhoff, 1977). Research into calibration of knowledge monitoring has largely involved the use of simple knowledge monitoring assessments like that developed by Tobias and Everson (2002). Knowledge monitoring assessments typically require the participant to judge whether and item, for example the meaning of a word, is known or not known, and then to either recall or recognize the correct meaning of the word. One adaptation of the format for the knowledge monitoring assessments used in prior research (e.g., Hartwig et al., 2012; Isaacson & 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 should allow an individual to successfully identify items of which they know the meanings and which items are not known. It is worth noting that for the purposes of the knowledge monitoring assessment, the number of items responded to correctly is not directly relevant. Rather than relying on the proportion of correct responses the results of the knowledge monitoring assessment 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 knowledge monitoring assessment (KMA) are generally presented in the form of a 2x2 contingency table like 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. Several methods are used to analyze the results of the knowledge monitoring assessment regarding calibration of knowledge monitoring.

Typically, to interpret the results of the knowledge monitoring assessment, a non-parametric correlation coefficient – Gamma – developed by Goodman and Kruskal (1954) has been calculated (e.g., Hartwig et al., 2012; Isaacson & Was, 2010). 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). It is important to note that 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 (2013) 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 which are not revealed in calculating gamma (Schraw et al., in press). 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 with 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, several variables are used to predict a binary outcome. In knowledge monitoring, an individual may make use of more than one 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 measurements of metacognitive knowledge-monitoring, then they each should be able to account for unique variance in academic achievement. That is, they should make contributions independent from one another if both constructs are important in the prediction of academic achievement. It is also likely that individual differences exist in each that contribute to learning. 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 a knowledge monitoring assessment 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 unique predictors of academic achievement. If sensitivity and specificity do serve as independent 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 the unique contributions of sensitivity and specificity to academic performance.

Method

Participants

Undergraduate students enrolled in several sections of an educational psychology course (N = 384) at a Midwestern university from the USA participated in the study in exchange for partial fulfilment of course requirements. All sections of the course were taught by the same instructor. All participants were of sophomore or junior class standing. Females made up 74.5% of the sample. 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.

Instruments

Knowledge monitoring assessment. The measure used to assess participants' accuracy of knowledge monitoring was adapted from Tobias and Everson (1995). The measure used in the present study involved presenting 50 vocabulary words to participants (33 related to the course, 17 general) one at a time. On the first presentation participants were to indicate whether 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, participants were given a multiple choice (5 possible responses) test on these same vocabulary words. Participants were required to complete this assessment online within the first two weeks of the course. Participants were informed that the knowledge monitoring assessment would not be graded and would have no effect on their grade in the course and asked to perform the assessment without assistance of any outside resources (e.g., the internet, a dictionary).

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, sensitivity and specificity were calculated for each individual participant. We also calculated Gamma to have a general sense of participants knowledge monitoring accuracy.

Final exam. For the purposes of the present study, we operationalized 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. Participants were allowed as much time as necessary to complete the exam.

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. To fulfil course requirements, participants 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. All ethical

standards of the American Psychological Association were followed in the treatment of participants and collection of the data.

Data Analysis

All further analysis 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). Data were analyzed using SPSS 24.

Table 1. *Example 2x2 contingency table for Knowledge monitoring assessment results*

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

Note. Gamma = (AD-BC)/(AD+BC); Sensitivity = A/(A+C); Specificity = D/(B+D)

Results

Table 2 presents the means and standard deviations for each of the measures of knowledge monitoring accuracy, the correctly identified items on the *Knowledge monitoring assessment,* and the percent correct on the *Final Exam*.

Table 2. *Means and Standard deviations*

Variable	*Mean*	*SD*
Gamma	0.58	0.25
Sensitivity	0.65	0.17
Specificity	0.65	0.16
Final Score	73.37	12.09

First, to confirm that gamma was related to academic achievement in the present sample we calculated a Pearson correlation between Gamma and the score (number correct) on the *Final Exam.* The correlation was significant ($r = .26$, $p < .001$), suggesting that gamma accounts for 6.8% of the variance in a *Final exam* taken 13 weeks after the *Knowledge monitoring assessment*.

To examine the relationships between sensitivity, specificity, and academic performance a multiple regression analysis was conducted. Results of the analysis are presented in Table 3.

Table 3. *Regression analysis predicting final exam performance from measures of metacognitive knowledge monitoring calibration sensitivity and specificity first* ($N = 361$)

Variable	*B*	*SE B*	*β*
Sensitivity	24.12	4.26	.34***
Specificity	6.24	4.49	.08
R^2		0.12	
F		18.28***	

Note. ***$p < .001$.

The model with sensitivity and specificity as predictors accounted for 8.7% of the variance in *Final Exam* performance, $F_{(2, 361)} = 18.23$, $p < .001$, $R^2Adj = .087$. The amount of variance accounted for is impressive considering the *Knowledge monitoring assessment* was completed online at least 13 weeks prior to the final exam, and also when considering how many factors impact test performance in an undergraduate level course. Of the 8.7% of the variance accounted for in *Final Exam* performance 8.06% was unique to sensitivity ($\beta = .34$, $t = 5.67$, $p < .001$), and only 0.49% was unique to specificity ($\beta = .08$, $t = 1.39$, $p > .05$).

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. In the current study, sensitivity (the ability to accurately judge what one knows) was a significant predictor of performance on a final exam 13 weeks prior to the exam. While the three variables should not be included simultaneously in analysis, it is recommended to report sensitivity and specificity alongside gamma rather than simply casting gamma 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 not 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 based on the results of this study alone. Further research may clarify this issue.

The results presented here suggest that sensitivity accounts for more variance in exam performance than that explained by gamma. At least in the case of using knowledge monitoring calibration to predict academic achievement it would seem that sensitivity is the preferred measure 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 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 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.

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.

The current results bring into question the relationships between knowledge monitoring, as measured by gamma, and academic achievement. Specifically, when effective knowledge monitoring predicted academic achievement as in Hartwig et al. (2012). Would the results of the Hartwig study been stronger or more detailed had they used sensitivity and specificity as there measures of knowledge monitoring rather than just gamma? Also, important to consider are those studies that suggest it may be possible to teach students to better monitor their knowledge (e.g., Al-Harthy et al., 2015; Isaacson & Was, 2010). Perhaps those studies that failed to find improved knowledge monitoring (e.g., Foster et al., 2019), would have been successful had they only focused on improving sensitivity.

Although the current study has important implications for both theory and practice, there are some limitations. First, the sample used in this study were undergraduates enrolled in large section of introductory course. Both the sample and the setting may limit the generalizability of the results. For example, students with less academic experience (i.e., K-12 students) may not have the same quality of knowledge monitoring as undergraduates. Experience with the material may also lead to different results. The course in which the study took place was an introductory course in education. Students in more advanced classes may also have more advance knowledge monitoring skill because they have more experience with the content. Another limitation may be the outcome measure. Using a multiple-choice final exam, as we did here, may produce different results than had we used an essay or exam or written assignment as our measure of knowledge. Both limitations mentioned provide impetus for future research.

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

Al-Harthy, I. S., Was, C. A., & Hassan, A. S. (2015). Poor performers are poor predictors of performance and they know it: Can they improve their prediction accuracy. *Journal of Global Research in Education and Social Science, 4*, 93 - 100.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*, 6 - 25. http://dx.doi.org/10.1080/00461520.2011.538645

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*, 1 - 19. http://dx.doi.org/10.1007/s11409-016-9158-6

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732 - 764. https://doi.org/10.1080/01621459.1954.10501231

Hartwig, M., Was, C. A., 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.2044-827 9.2011.02038.x

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

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159 - 183. https://doi.org/10.1016/0030-5073(77)9000 1-0

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. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*, 982 - 990. 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*, 321 - 332. doi: 10.1002/acp.2350090405

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). 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.