Student Metacognitive Monitoring:

Predicting Test Achievement from Judgment Accuracy

Alfred Valdez[1]

[1] College of Education, New Mexico State University, New Mexico, USA

Correspondence: Alfred Valdez, College of Education, Special Education/Communication Disorders Department, MSC 3SPE, New Mexico State University, Las Cruces New Mexico, 88003-8001, USA. Tel: 1-575-646-7607. E-mail: valdez1@nmsu.edu

Received: April 6, 2013 Accepted: April 30, 2013 Online Published: May 3, 2013

doi:10.5430/ijhe.v2n2p141 URL: http://dx.doi.org/10.5430/ijhe.v2n2p141

Abstract

Metacognitive monitoring processes have been shown to be critical determinants of human learning. Metacognitive monitoring consist of various knowledge estimates that enable learners to engage in self-regulatory processes important for both the acquisition of knowledge and the monitoring of one's knowledge when engaged in assessment. This study investigated the reliability and the predictive validity of one measure of metacognitive monitoring, *absolute accuracy*. *Absolute accuracy* pertains to how accurately students judge their knowledge relative to their actual performance on a test. Students show poor *absolute accuracy* if they judge their knowledge as correct when in fact it is in error, or if they judge their knowledge in error when in fact it is correct. Contrary to previous studies, this study revealed that *absolute accuracy* indices may be reliably measured in an authentic setting. The study also showed that *absolute accuracy* was significantly correlated with concurrent and final exam performance.

Keywords: Calibration, Metacognition, Absolute accuracy

1. Introduction

Many would agree that self-regulated learners are not only better prepared to engage in the learning process but that they approach the task of learning with greater efficiency (Butler & Winne, 1995; Carver & Scheier, 1998; Schunk, 1996). Self-regulated learners deploy a number of cognitive, affective, motivational, and metacognitive processes that support their learning (Schraw, Crippen, & Hartley, 2006). This investigation will focus on one aspect of metacognitive monitoring processes, calibration, and its influence on test performance.

Metacognition consists of monitoring processes (Nelson & Narens, 1990) and control processes (McCormick, 2003; Nelson & Narens, 1990; Schraw & Moshman, 1995). Metacognitive monitoring primarily pertains to learners' estimates of their own knowledge, that is, learners' knowledge of strategies that support cognition and their knowledge of conditions that dictate when and how to execute strategies that might influence their own learning. Metacognitive control, on the other hand, pertains to a learner's ability to select, monitor and regulate their knowledge to best support their own learning.

In test (and test-like) situations, students' confidence estimates could conceivably influence their ability to accurately control and adjust their responses to test items. For example, students who are erroneously overconfident about their content knowledge may more easily be drawn to multiple-choice selections that are near approximations of the correct answer. In addition, learners who are erroneously overconfident may fail to execute test-taking strategies (e.g., comprehension monitoring of test questions, identification and selection of easier test questions first, use of question stems to eliminate and select best responses, etc.) that help to guide the student toward the correct answer choice (test-taking strategies see Cohen & Upton, 2007; Dodeen, 2008; Eunsook, Sas, & Sas, 2006).

Conversely, students who are erroneously underconfident may spend unnecessary time on a few test items and thereby limit their opportunity to execute test-taking strategies. These students may also experience the possible detrimental effects that self-doubt plays on their own self-efficacy beliefs.

Students' ability to accurately match their confidence of success with the actual outcome of a test has been described as calibration (Pieschl, 2009; Stone, 2000; Winne, 2004). That is, if a learner judges their confidence of success for a given test at 80% correct and the outcome of that test is 80% correct, then the learner has exhibited perfect

calibration. That same 80% correct judgment given a 70% correct test outcome would be an example of *overconfidence* and for a 90% correct test outcome an example of *underconfidence*.

While self-regulated learning is best described as a cyclical self-correcting process, calibration is more aptly described as a learner's self-reflection on that process. What is less clear is how this manner of self-reflection (i.e., calibration) influences self-regulated learning during an actual test setting. One might assume that well calibrated test-takers (i.e., those better able to monitor their knowledge while taking a test) outperform less calibrated learners. However, there is little empirical evidence to support this proposition (Stone, 2000).

1.1 Metacognitive Monitoring

Arriving at a confidence judgment pertaining to one's own knowledge is one of many metacognitive monitoring processes (Nelson & Narens, 1990; Nelson, 1999). But how important is that process when arriving at a correct answer choice? Consider the test-question, "What is the capitol of New Mexico?" Consider also that the retrieved response to that question is an incorrect response, "Mexico City", (the *cognitive* element) and that the learner reports an extremely high confidence level, "100% confident", that their answer was correct (the *metacognitive* element). In this example both the cognitive (Mexico City) and the metacognitive (100% confidence) elements are erroneous. That is, the information retrieved from memory (Mexico City) does not correspond with the correct answer (Santa Fe), and the metacognitive judgment (100% certainty) is directly opposed to the test outcome (incorrect response). This is a case where metacognitive monitoring failed to detect, consider and correct an initial incorrect response. More importantly, this may be an instance where the failure to monitor one's knowledge draws a learner toward an erroneous response to the test question.

Gaskins, Dunn, Forte, Wood, and Riley (1996) found that students who express a high level of certainty in their test answers did not utilize strategies that might support the further consideration, and the possible correction, of their answer. It appears that students must feel a certain degree of uncertainty; otherwise, they are less likely to engage in metacognitive control strategies such as moving on to another question to conserve on allotted test-taking time, or using a prior-learned mnemonic device to more effectively retrieve the correct answer (Flannelly, 2001). Less is known about students who are uncertain about their test knowledge.

Although a considerable body of literature exists on the topic of calibration, the majority of findings have been obtained from laboratory studies that utilized contrived learning measures such as paragraph comprehension or paired associate learning (Pieschl, 2009). These studies have provided an important foundation for the study of this construct, but more current reviews (Hacker, Bol, & Keener, 2008; McCormick, 2003) have suggested that it is time to investigate the construct of calibration in more authentic settings such as classrooms and to use authentic performance measures such as course grades, grades on tests and grades on course assignments as outcome measures.

1.2 Purpose

This investigation will address two issues pertaining to the measurement of *calibration*. First, given that few investigations have directly addressed the psychometric properties of various measures of calibration (Schraw, 2009; Thompson, 1998, 1999) this study will investigate the internal structure reliability (i.e., coefficient alpha) of one measure of calibration, *absolute accuracy*, for students enrolled in a post-secondary course in language development. Second, this study will investigate how well *absolute accuracy* estimates predict concurrent and final exam performance for those same post-secondary students. Few studies have examined how well *absolute accuracy* estimates predict concurrent and later performance on authentic student measures such as course quizzes and exams. A strong inverse relationship between estimates of *absolute accuracy* and the concurrent test result would suggest that students who are more accurate at judging their test knowledge while taking a test perform better on the concurrent and later measure.

2. Methods

2.1 Participants

Twenty-four students (3 male and 21 female) enrolled in an undergraduate course (Language Acquisition) in the Communication Disorders Program were included in this study. The mean age for all students was 20 years. One student was over 50 years of age.

2.2 Procedure

During each of six multiple-choice quizzes and on the cumulative final exam, students were asked to rate each quiz item on the probability that their answer was correct by circling one of four certainty estimates (0%, 33%, 66%, and 100% certain). Prior to the quiz, the instructor explained the confidence scale by stating, "If you are absolutely sure that your answer is correct, rate your answer 100%. If you are absolutely unsure that your answer is correct, rate your answer 0%. If you are mostly certain that your answer is correct, rate your answer 66%, but if you are mostly uncertain that your answer is correct, rate your answer 33%". Students were asked to rate their answers immediately after they decided on a final response to each question. Each of the six quizzes contained 21 questions and the final exam contained 51 questions.

2.3 Data Analysis

2.3.1 Absolute Accuracy

Schraw's (2009) formula for *absolute accuracy* was used to calculate meta-cognitive judgment accuracy. Schraw's formula compares the students' confidence estimates with their actual performance. A confidence estimate of 0% when an item is shown to be in error or a confidence estimate of 100% when an item is shown to be correct would result in a discrepancy score of zero (i.e., perfect calibration). Any mismatch of confidence judgment and score outcome (i.e., correct/incorrect) would result in a discrepancy score that indicates either overconfidence (i.e., some level of confidence that the item was correct but in fact the item was in error) or underconfidence (i.e., some level of confidence that the item was in error when in fact the item was correct). Each discrepancy estimate is squared, the squared deviations are summed and the mean deviation score represents the *absolute accuracy* estimate for each student. *Absolute accuracy* estimates closest to zero indicate a perfect match between a student's judgment and their actual performance. Scores greater than zero indicate greater degrees of judgment inaccuracy.

2.3.2 Reliability

Reliability is a necessary but insufficient condition for score validity. In its broadest sense, reliability is conceived of as score consistency. Score reliability was estimated using an internal consistency approach, coefficient alpha. Coefficient alpha was chosen because of the known practical problems inherent with test-retest and equivalent forms approaches of estimating score reliability (see Pedhazur & Schmelkin, 1991). Conceptually, coefficient alpha treats each test item as a parallel (or equivalent) form of the measure of interest. Therefore, a high degree of consistency in measurement among the test items offers evidence of high score reliability.

The computational elements for coefficient alpha are the number of test items (k), the sum of the variances for each test item (sum_var), and the sum of all pairwise covariances for each of the test items (sum_covar). The formula for coefficient alpha, Alpha = k/(k-1) * [1 – (sum_var) / (sum_var) + 2(sum_cov)] is taken from Pedhazur & Schmelkin (1991, pp. 92-97). Covariance is a measure of correspondence between two variables and total item variance may be conceived of as the precision of the composite score of a given measure. The value derived from the k/(k-1) portion of the formula will typically approximate unity (the value one), particularly for measures with many (more than 9) items and will therefore have little influence on the coefficient derived from this formula.

As shown in the formula described earlier, the relationship of total variance and item covariance suggests that a highly precise measure (small total variance) combined with a measure with high test item correspondence (or cohesive internal structure) will result in a small value. This small value, when subtracted from 1, will result in a large alpha coefficient or an alpha coefficient that implies high score reliability. This formula suggests that coefficient alpha takes into consideration the precision of the score that represents the composite measure in relation to the homogeneity of the items that make up the composite score. In other words, test items showing a high level of inter-item covariance and a low level of total score variance will generally show a large coefficient alpha.

3. Results

3.1 Reliability

Chronbach's alpha reliability coefficients were used to assess internal structure reliability on *absolute accuracy* estimates for each of the six quizzes. Schraw's (2009) *absolute accuracy* estimates take the average, or aggregate, of all confidence judgments within one test. Therefore, Chronbach's alpha was calculated using these six aggregate values as the unit of analysis to arrive at the internal consistency estimate of reliability. The internal consistency estimate of reliability was .81 indicating adequate reliability (Nunnally & Bernstein, 1994) for the six *absolute accuracy* estimates.

Table 1. Absolute accuracy and test score correlations for the six quizzes and the final exam

	Abs Quiz 1	Abs Quiz 2	Abs Quiz 3	Abs Quiz 4	Abs Quiz 5	Abs Quiz 6	Abs Final Exam
Mean							
Abs Est. [a]	.151	.131	.183	.170	.176	.241	.191
(SD)	(.062)	(.096)	(.086)	(.066)	(.093)	(.094)	(.087)
Quiz 1	**-.75***						
Quiz 2		**-.72***					
Quiz 3			**-.76***				
Quiz 4				**-.62***			
Quiz 5					**-.70***		
Quiz 6						**-.67***	
Final Exam	**-.61***	**-.71***	**-.33**	**-.68***	**-.58***	**.64***	**-.92***

Note:

Abs Est. = *Absolute Accuracy* Estimate

* denotes statistically significant correlation (alpha < .05)

Test score and *absolute accuracy* estimate concurrent correlations are on the diagonal.

Absolute accuracy estimates and final exam score predictive correlations are along the bottom row.

Absolute accuracy means and standard deviations for each of the six quizzes and the final exam score are along the top row.

3.2 Absolute Accuracy Descriptive Findings

The top row of Table 1 shows the means and standard deviations of the *absolute accuracy* estimates for each of the six quizzes and for the final exam. *Absolute accuracy estimates* are an index of judgment precision. Mean values closer to zero indicate more precise average judgments made by students. Students were most precise in their judgments on quiz two, followed by quiz one, quiz four, quiz five, quiz three, and quiz six. Students' judgment precision ranged from .13 to .24 across the six tests. In general, early judgment precision was better than later judgment precision.

3.3 Prediction of Concurrent Test Performance

Table 1 also shows how well *absolute accuracy* estimates predict concurrent test performance. As shown by the correlation coefficients on the diagonal, all seven *absolute accuracy* estimates significantly predicted concurrent test performance on each of the six quizzes. The extent to which the seven *absolute accuracy* estimates predicted concurrent test performance was greatest on the final exam with 85 percent of the variation in final exam scores predicted by concurrent *absolute accuracy* estimates. Concurrent predictions of *absolute accuracy* estimates on quiz performance was somewhat less with 56, 52, 58, 38, 49 and 45 percent of the variation in quiz performance (quizzes one through six respectively) predicted by concurrent *absolute accuracy* indices.

3.4 Prediction of Final Test Performance

Table 1 also shows how well *absolute accuracy* estimates predict final exam performance. As shown by the correlation coefficients on the last line of Table 1, five of the six *absolute accuracy* quiz estimates significantly predicted final exam performance. The extent to which each *absolute accuracy* estimate predicted final exam performance was greatest on quiz two (r = .71) with fifty percent of the variation in final exam scores predicted by the quiz two *absolute accuracy* estimates. Predictions of final exam scores based on quiz *absolute accuracy* estimates were similar for quizzes one, four, five and six (r = .61, .68, .58, and .64 respectively). *Absolute accuracy* indices for quiz three (r = .33) did not significantly predict final exam performance.

4. Discussion

Metacognitive monitoring may be a novel term for some, but it is not a novel idea. Most individuals know how to deal with indoor room temperature. Locate the device that controls the room temperature, such as the thermostat, and adjust it lower if it is too hot in the room or adjust it higher if it is too cold in a room. Metacognitive monitoring is similar in concept in that it relies first on the detection of a mismatch between an intended outcome and the actual outcome (Carver & Scheier, 1998). Just as an erroneously calibrated thermostat will fail to appropriately adjust room temperature, a poorly calibrated learner may fail to monitor, select, and enact appropriate strategies in order to perform well in a testing situation.

This investigation proposed that the ability to accurately monitor one's knowledge during test-taking, *absolute accuracy,* strongly accounts for the variation in one's test performance. That is, during a test, learners who are more accurate at judging their test responses will perform better on the test they are currently taking. However, in order to investigate the relationship between *absolute accuracy* estimates and test performance, one must first determine the reliability of those estimates.

This study first addressed the issue of score reliability of *absolute accuracy,* and showed that *absolute accuracy* appears to reliably measure on-line metacognitive monitoring. Internal structure reliability for the aggregate estimates of *absolute accuracy* were shown to be good (r = 81). Score reliability was much better than reliabilities reported elsewhere (Thompson, 1999; Veenman et al., 2006). This may be due to the fact that reliability estimates for this study utilized a wide range of item difficulty (item difficulty ranged from .30 to .90) and item-by-item confidence judgments. Schraw (2009) has suggested that these factors should improve estimates of score reliability.

The second question concerns the relationship between *absolute accuracy* estimates and test performance. This investigation found significant and moderate to large correlations between students' *absolute accurate* estimates and their concurrent and final exam test performance. Students who were more accurate at judging their responses scored higher on tests than students who were less accurate at judging their responses. While this finding supports the importance of students accurately judging their test responses, it raises other questions. For example, what kinds of test-taking behaviors emerge as a result of judgment inaccuracies?

Student *absolute accuracy* estimates appear to be a reliable measure of metacognitive monitoring and an important predictor of students' success on authentic measures of classroom learning. While these results suggest that *absolute accuracy* offers a reliable and important index of metacognitive monitoring, these findings are limited to a specific group of post-secondary students, and a specific topic of instruction. This finding raises questions pertaining to judgment accuracy and learning. Future studies should investigate those factors that might support or fail to support judgment accuracy. For example, what test situation variables (i.e., learner preparation or teacher support) encourage improved metacognitive monitoring? How does judgment bias (i.e., overconfidence versus underconfidence) support the choice of test-taking strategies? Finally, students making confidence judgments during a test may itself lead to inaccurate response judgment, inaccurate response choice, or unwillingness to record a confidence judgment. It would be useful to determine qualitatively how students eventually arrive at their judgments of response accuracy.

References

Butler, D. L. & Winne, P. H. (1995). Feedback and self-regulated learning. *Review of Educational Research, 65,* 245-281. http://dx.doi.org/10.3102/00346543065003245

Carver, C. S. & Scheier, M. F. (1998). *On the self-regulation of behavior*. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174794

Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*(2), 209-250. http://dx.doi.org/10.1177/0265532207076364

Dodeen, H. (2008). Assessing test-taking strategies of university students: developing a scale and estimating its psychometric indices. *Assessment & Evaluation In Higher Education, 33*(4), 409-419. http://dx.doi.org/10.1080/02602930701562874

Eunsook, H., Sas, M., & Sas, J. C. (2006). Test-Taking Strategies of High and Low Mathematics Achievers. *Journal of Educational Research, 99* (3), 144-155. http://dx.doi.org/10.3200/JOER.99.3.144-155

Flannelly, L. (2001). Using feedback to reduce students' judgment bias on test questions. J*ournal of Nursing Education, 40* (1), 10-16. PMid:11198904

Gaskins, S., Dunn, L., Forte, L., Wood, F., & Riley, P. (1996). Student perceptions of changing answers on multiple choice examinations. *Journal of Nursing Education, 35*(2), 88-90. PMid:8926526

Hacker, D.J., Bol, L, & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of metamemory and memory*, (pp. 429-456). Hillsdale, NJ: Psychology Press.

McCormick, C. B. (2003). Metacognition and learning. In W. B. Reynolds & G. E. Miller (Eds.), *Handbook of Psychology: Educational psychology* (pp. 79-102). Hoboken, NY: Wiley.

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

Nelson, T. O. (1999). Cognition versus metacognition. In R. J. Sternberg (Ed.) *The nature of cognition* (pp. 625-641). Cambridge, MA: MIT Press.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory (3rd ed.).* New York: McGraw-Hill.

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement design and analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition & Learning, 4(1),* 3-31. http://dx.doi.org/10.1007/s11409-008-9030-4

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition & Learning, 4(1),* 33-45. http://dx.doi.org/10.1007/s11409-008-9031-3

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education, 36* (1-2), 111-139. http://dx.doi.org/10.1007/s11165-005-3917-8

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7* (4), 351-374. http://dx.doi.org/10.1007/BF02212307

Schunk, D. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal, 33*(2), 359–382. http://dx.doi.org/10.3102/00028312033002359

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. http://dx.doi.org/10.1023/A:1009084430926

Thompson, W. (1998). Metamemory accuracy: Effects of feedback and the stability of individual differences. *American Journal of Psychology, 111*(1), 33-42. http://dx.doi.org/10.2307/1423535

Thompson, W. B. (1999). Individual differences in memory monitoring accuracy. *Learning and Individual Differences, 11*, 365–376. http://dx.doi.org/10.1016/S1041-6080(99)80009-0

Veenman, M. V. J., Van Hout-Walters, B. H. A., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. http://dx.doi.org/10.1007/s11409-006-6893-0

Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*(6), 466-488. http://dx.doi.org/10.1016/j.ijer.2005.08.012