Learning and Instruction 47 (2017) 33–42

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

The measurement of calibration in real contexts

Teomara Rutherford

Department of Teacher Education and Learning Sciences, College of Education, North Carolina State University, Raleigh, NC 27695, United States

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form
12 October 2016
Accepted 17 October 2016
Available online 28 October 2016

Keywords:
Metacognitive monitoring
Accuracy
Calibration
Data
Self-regulated learning

ABSTRACT

Accurate judgment of performance, or calibration, is an important element of self-regulated learning (SRL) and itself has been an area of growing study. The current study contributes to work on calibration by presenting practical and predictive results of varying calibration measures from authentic educational data: elementary-aged students' interactions with a year-long digital mathematics curriculum. Comparison of predictive validity of measures show only small differences in explained variance in models predicting posttest performance while controlling for pretest. A combined model including Sensitivity and Specificity outperforms other single measures, confirming results in Schraw, Kuch, & Gutierrez (2013); however, results show that student patterns of calibration within these data differ from those assumed in simulation studies and these differences have implications for the calculability of popular calibration measures.

1. Introduction

The ability to accurately judge one's performance is a foundational aspect of self-regulated learning (see Winne & Hadwin, 1998; Zimmerman, 2008). This accuracy is sometimes termed *calibration*, and across numerous contexts has been found to have its own relation with academic achievement (Stone, 2000; e.g., Bol, Riggs, Hacker, Dickerson, & Nunnery, 2010; Ots, 2013; Koku & Qureshi, 2004). Although calibration is noted as an important skill for learning (Alexander, 2013), and is a popular area of research, unanswered questions remain about the nature of calibration. The current study contributes to extant discussions regarding the best way to measure and calculate calibration (e.g., Masson & Rotello, 2009; Nietfeld, Enders, & Schraw, 2006; Schraw, 2009) by presenting a comparison of the practical and predictive results of varying calibration calculations for data obtained from student interactions with a year-long digital mathematics curriculum. In addition, it contributes to the empirical research regarding calibration by presenting results on the predictive power of calibration measures in an authentic elementary mathematics setting.

Calibration generally refers to the agreement between perception of task performance and actual performance (Nietfeld et al., 2006; Stone, 2000) and can be operationalized in a variety of ways. In particular, the choice of how to calculate measures of calibration affects conclusions drawn. Researchers can focus on absolute calibration (e.g., Bol & Hacker, 2001; Huff & Nietfield, 2009) or can investigate the direction of the calibration (e.g., Chen, 2002; Mengelkamp & Bannert, 2010). By way of illustration, in Fig. 1, two students, Sarah and Jenny, have the same level of calibration (looking item-by-item at the agreement between confidence and correctness), but Sarah displays an overconfident bias whereas Jenny is not consistently biased in either direction. Even among measures only looking at agreement, calculations may differ in how they treat these agreements. It is differences between these calculations of calibration upon which this study focuses, asking (1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments? and (2) Among these measures, which display the greatest predictive validity?

1.1. Calibration's role in self-regulated learning

Although there are a number of theories of self-regulated learning, all generally involve some process in which students set goals, monitor their progress toward these goals, and adjust their performance accordingly (e.g., Pintrich, 2000; Zimmerman, 1989; Winne, 1995). Student ability to accurately assess performance is important at each stage of the process. In the planning or forethought phase, student self-efficacy informs goal-setting (Bandura, 1986). Although slightly over-positive self-efficacy may be most adaptive for setting attainable goals, an over-inflated sense of self-efficacy may result setting too lofty a goal, resulting in failure, accompanied by discouragement and disengagement (Bandura,

E-mail address: teyarutherford@gmail.com.

http://dx.doi.org/10.1016/j.learninstruc.2016.10.006
0959-4752/

Sarah

#	Acc	Conf	Match
1	Y	↑	☑
2	N	↑	☒
3	N	↑	☒
4	N	↓	☑
5	N	↓	☑
	20%	60%	60%

Jenny

#	Acc	Conf	Match
1	Y	↑	☑
2	N	↑	☒
3	Y	↑	☑
4	Y	↓	☒
5	N	↓	☑
	60%	60%	60%

Fig. 1. Illustration of calculation of item-by-item as compared to more macro levels of calibration.

1986; Winne, 2004). As students work toward their goals, they adjust their strategies and resource allocation as they monitor their success (Nelson, 1996; Pintrich, 2004; Winne, 2001). Those students who determine that they are not performing at an appropriate level will attempt to rectify the situation by exercising control (Winne, 1995). It is this determination where the current study is focused: measures of calibration provide us with indications of student ability to accurately assess their performance.

1.2. Comparisons of calibration measures

In selecting measures of calibration, prior research has noted the importance of aligning the purpose of the study with the selected measure (Boekaerts & Rozendaal, 2010; Nietfeld et al., 2006; Schraw, 2009). Various measures may be complementary in that they can provide information on absolute accuracy, bias, or the ability to distinguish between correct and incorrect items (see Boekaerts & Rozendaal, 2010; Schraw, 2009). Choice of measures also can be driven by underlying assumptions about the monitoring process, for example, whether monitoring of potential correct and incorrect answers happens through a single process or separately through distinct processes (Schraw, Kuch, & Gutierrez, 2013); see discussion 1.2.4, below).

Practical considerations beyond the match with research question may also guide the choice of measure. For example, it has been suggested that for young children, measures with fewer choices reduce the cognitive load and allow for more accurate calibration scores (see e.g, Lyons & Ghetti, 2011). The Jenny and Sarah example illustrates a simplified dichotomous measure wherein students indicate whether they feel confident or not confident for each answer given.

1.2.1. The calculation of calibration indices and the practical issue of missing quadrants

The use of such a dichotomous measure in relating accuracy to judgments of confidence results in a 2 × 2 contingency table with cells depicted in Fig. 2. Looking to our examples: of the five quiz questions, Sarah would have one question in cell A, two each in cells B and D, and none in cell C. Jenny would have two questions in cell A and one each in the other three cells. Numerous indices have been created for the calculation of agreement based on the contents of these cells (see Feuerman & Miller, 2008; Schraw, 2009; Schraw et al., 2013). Table 1 presents a number of common indices expressed as functions of cells A through D and largely draws on descriptions of these formulas presented in Schraw et al. (2013) work. Some have emerged as more popular that others: Gamma

A. Confident & Correct	B. Confident & Incorrect
C. Not Confident & Correct	D. Not Confident & Incorrect

Fig. 2. 2 × 2 contingency table expressing the relations between accuracy and confidence.

(e.g., Mengelkamp & Bannert, 2010; Thiede, Anderson, & Therriault, 2003), d' or discrimination (e.g., Boekaerts & Rozendaal, 2010; Macmillan & Creelman, 1996), and G Index (e.g., Schraw, 1995; Tobias & Everson, 1998) have been particularly popular within metacognition and self-regulated learning research. Sensitivity and specificity have been more popular in medical research, where they represent successful detection of the presence or absence of a condition, respectively (e.g., Warnick, Bracken, & Kasl, 2008). Schraw et al. (2013) divided these ten common indices into interpretive families based on the dimensions purportedly captured by each measure—families are specified in Table 1. These and other measures have other empirical justifications (e.g., Gamma may be most useful in determining consistency of judgments whereas G Index may be most useful in measuring changes in calibration, see Nietfeld et al., 2006), and there are also practical ramifications of the selection of one measure over another. Due to the nature of the

Table 1
Common measures of calibration from 2 × 2 contingency tables.

Index	Formula
Sensitivity[a]	A/(A + C)
Specificity[a]	D/(B + D)
Simple Matching[b]	(A + D)/(A + B + C + D)
G Index or Hamann coefficient[b]	(A + D) − (B + C)/(A + B + C + D)
Odds Ratio[c]	AD/BC
Goodman-Kruskal Gamma[c]	(AD − BD)/(AD + BC)
Kappa[c]	2*(AD − BC)/[(A + B)(B + D) + (A + C)(C + D)]
Phi[c]	(AD − BC)/[(A + B)(B + D)(A + C)(C + D)]$^{1/2}$
Sokal Reverse[d]	[1 − [(A + D)/(A + B + C + D)]]$^{1/2}$
Discrimination (d')[e]	z(A/(A + C)) − z(B/(B + D))

Note. Formulas as represented in Schraw et al. (2013). Superscripts indicate the category of measurement as defined by Schraw et al. (2013): (a) Diagnostic efficiency, (b) Agreement, (c) Association, (d) Binary distance, and (e) Discrimination.

formula calculations, the distribution of data within 2 × 2 contingency tables affects each of the measures differently. For some, the lack of selections that fall in certain quadrants is especially problematic. As an example, Gamma is undefined when certain combinations of quadrants are missing (A and C, A and B, D and C, D and B) and can be heavily distorted when even one quadrant is zero (see Kuch, 2012).

1.2.2. Studies of simulated data

These distortions from missing quadrants have been quantified and discussed in prior research (e.g., Kuch, 2012; Masson & Rotello, 2009; Nietfeld et al., 2006; Schraw, Kuch, & Roberts, 2011). In particular, previous work has examined the extent of distortion due to number of test questions and difficulty of questions (e.g., Kuch, 2012; Schraw et al., 2011), the comparative distortion between measures (e.g., Kuch, 2012; Nietfeld et al., 2006), and solutions for eliminating this distortion (e.g., Hautus, 1995; Miller, 1996; Schraw et al., 2011). The bulk of this research is conducted by examining the behavior of the measures using simulated data. A typical process uses a Monte Carlo simulation to create responses to tests of lengths from six to 1000 questions (often assuming the correct/non-distorted estimates will be present at 1000 questions). To simulate a distribution of responses due to chance, each question response is randomly assigned to one of the four quadrants resulting in approximately 25% of the responses in each cell. To simulate a moderately accurate condition, which is often assumed to approximate real-life (see Kuch, 2012; Nietfeld et al., 2006; Schraw et al., 2013), 50% of the responses are assigned to cell A (confident and correct) and then the remaining 50% are randomly assigned across all four cells. This results in a distribution of 62.5% in cell A and 12.5% in each of the other cells. Based on such simulated datasets, Nietfeld et al. (2006) concluded that G Index was more reliable across varying test sizes than Gamma, a conclusion supported by Kuch (2012). Schraw et al. (2011) suggested that for Gamma to be reliable, it needed to be calculated from moderately difficult tests of at least 20 questions. This suggestion for test size was based on data distribution that simulated an equal likelihood of being in cells B through D, even in conditions meant to approximate tests of moderate difficulty. However, theory surrounding metacognitive judgments does not support this distribution of responses. As difficulty increases, it is more likely that test-takers will make *over*confident judgments (see Kruger & Dunning, 1999), likely resulting in a paucity of quadrant D responses, even when test-takers make more incorrect responses.

1.2.3. Solutions to missing quadrants

To attempt to avoid distortion from zero quadrants, two tactics have been previously used. A number, such as 0.05, can be added either to only the missing quadrant or to all four quadrants (see Hautus, 1995). Schraw et al. (2011) and Miller (1996) demonstrated with simulated data of at least 1000 questions that adding a number to the missing quadrant did not eliminate data distortion and that distortion varied depending on the exact number added and on the value for calibration that would have been observed without. Hautus (1995) noted that although neither commonly-used convention for handling missing quadrants completely replicated true non-problematic data, the practice of adding a value across all quadrants came closer. Schraw et al. (2011) discouraged researchers from the practice of modifying the data, noting that modifying the instrument to a moderately difficult 25-question test should solve the problem of missing quadrants.

1.2.4. Relation and strength of measures

Schraw et al. (2013) used simulated data to move beyond the practical difficulties of distortion due to zero quadrants to concentrate on the interrelation of common measures. The authors classified each of the measures in Table 1 into one of five interpretive families based on the main purpose for each measure (diagnostic efficiency, agreement, association, binary distance, and discrimination) and noted that there were theoretical reasons for possible degrees of correlation between each of the measures. Using simulated datasets of 1000 question quizzes, they conducted confirmatory factor analyses to test three competing models of metacognitive monitoring: the first, based on the Nelson and Narens (1990) one-factor solution, the second, specifying a two-factor solution with Sensitivity and Specificity as orthogonal processes subsuming variance in all the other measures (see Feuerman & Miller, 2008), and the third, specifying five interrelated factors based on the theoretical families. The authors concluded that the second model was the best-fitting for their simulated data in both the chance and moderate accuracy conditions and that a combined model including both Sensitivity and Specificity, which are accurate judgements of certainty and uncertainty, respectively, would explain the most variance in metacognitive monitoring.

1.2.5. Moving beyond simulated data

Schraw et al. (2013) had theoretical reasons for supporting their advocacy of a combined Sensitivity/Specificity model; however, their conclusions were based off of simulated data and did not investigate the predictive validity of each of the measures studied. A comparison of these common measures of calibration has rarely been undertaken using non-simulated data, especially those from an authentic learning task (Schraw, 2009). Most studies of calibration with authentic learning tasks include relatively small samples using tests of limited size (Schraw et al., 2013; e.g., Huff & Nietfield, 2009; Pajares & Miller, 1997). This makes comparisons between measures difficult. Even those that have used multiple measures (e.g., Allwood, Jonsson, & Granhag, 2005; Boekaerts & Rozendaal, 2010) do not explore practical limitations of the data and often focus on calibration as an outcome, without examining differences in predictive validity between scores. One exception is Schraw and colleagues' extension of their 2013 study (Schraw, Kuch, Gutierrez, & Richmond, 2014) in which the authors used non-simulated data and replicated the prior finding that a combined Sensitivity/Specificity model explained more variance than any of the single-measure models. However, the study tasks used were decontextualized tests of vocabulary, computation, and spatial ability, and although the authors note data loss from the limitations of their 15-item tests, they do not explore the ramifications of this loss.

1.3. Individual differences and influence on calibration

As noted in Section 1.2.1, the calculability of calibration is influenced by item difficulty, test-taker performance, and number of questions attempted. These factors are themselves likely influenced by individual differences among test takers. Prior research has indicated that student gender and achievement level are related to accuracy of student calibration (e.g., Boekaerts & Rozendaal, 2010; Gutierrez & Price, 2016; Stone, 2000). Mathematics achievement itself is related to student background, such as socioeconomic status (e.g., Sirin, 2005). Moving beyond simulated data will allow for an exploration of how these individual differences influence proximal drivers of calculability.

1.4. Current contribution

The current paper utilizes data that address the shortcomings of both simulated data and authentic data as typically used. The data herein are authentic in that they come from student interactions

with learning materials administered as part of their normal mathematics classes, and the data are not subject to the typical limitations of real-world data in that the sample size is large (over 4000 students) and aggregated across the year's curriculum to produce a test with over 200 questions. These data permit an examination of how the actual distortion from zero quadrants affects the calculation of different calibration measures and how these measures, once calculated, differentially predict measures of achievement gains.

2. Method

2.1. Research design and population

The current study uses data from an evaluation of the MIND Research Institute's Spatial Temporal (ST) Math software. Within the larger project, the effectiveness of the ST Math digital mathematics curriculum was evaluated using a randomized control trial of 52 schools (see Rutherford et al., 2014). The project schools included two cohorts with a staggered implementation design. This manuscript concentrates on the 18 Cohort 2 schools, which began implementing ST Math in the 2009–2010 school year and for whom data within ST Math were collected and provided to the researchers. Participating students played the ST Math games 90 min each week during each school year starting in 2009–2010.

The overall ST Math study sample consisted of all second through fifth grade students (approximately aged six through eleven) in 52 low-performing schools within ten districts in Southern California. Each school enrolled 200 to 800 students in these grades in a given year. Descriptive statistics for the study sample are shown in Table 2. In comparison with the populations of the county and to the state of California as a whole (CADOE STAR, 2011), the study schools contained a larger percentage of Hispanic students (85%), English Language Learners (ELLs, 65%), and students eligible for free/reduced price lunch (81%). This study focuses on the 4281 students who were using ST Math in the 2010–2011 school year: approximately half of the second, third, and fifth graders in the Cohort 2 study schools, and all of the fourth graders.

Table 2
Comparison of sample descriptives to county and state.

	Study Sample	Count	County	California
	Mean/Percent		Mean/Percent	Mean/Percent
Male	52%	4147	50%	49%
Free/Reduced Lunch	81%	4147	46%	57%
Hispanic	85%	4147	47%	50%
White	8%	4147	31%	26%
Asian	3%	4147	14%	9%
Other Ethnicity	3%	4147	8%	23%
Eng Language Learner	65%	4146	39%	32%
N	4281		110,402	1,401,811

Note. Column 1 is calculated from available data within the sample. County and California data aggregated for grades two through four in 2008–2009 from the California STAR reporting website: http://star.cde.ca.gov/star2011.

2.2. Instruments/measures/sources of data

2.2.1. ST math quiz data

Within ST Math, students completed up to 24 mathematics objectives, depending on grade level. As students started a new objective module, they took a five to ten-question pretest on the content within that module and specified their confidence (sure or not sure) in each answer they gave (Fig. 3). After the module, they took a five to ten-question posttest, also selecting their confidence level. The combination of this accuracy and confidence data provided information on student calibration. MIND provided item-by-item quiz answers, accuracy, and confidence ratings for each student who engaged with the ST Math curriculum during the 2010–2011 school year. Of the quizzes students encountered, 90% were five-question quizzes. Each year included up to 48 of these quizzes (counting pre and post separately), depending on grade-level. Although content and difficulty of quizzes varied by objective, high within-student reliability of pretest accuracy was indicated by Cronbach's alpha of 0.87. Calibration measures were calculated from scores aggregated across all quizzes taken during the year's curriculum. For each student, quadrants A through D (see Fig. 2) were aggregated within quadrant, and formulas provided in Table 1 were applied to result in the ten measures of calibration. For question one, pretest and posttest question data were combined. For question two, only pretest calibration data were used. Together, pre and posttest quiz questions included up to 270 questions for second grade, 280 questions for third grade, 270 questions for fourth grade, and 230 questions for fifth grade. Number of total questions answered was also calculated for each student.

2.2.2. Demographics

Gender, ethnicity, free/reduced lunch, and ELL status were provided by the school districts. Ethnicity is represented in the analyses by five groups: Hispanic, Vietnamese, Black, White, and Other Ethnicity, to represent the largest ethnic groups within the sample. Reported English Language Learner (ELL) status was determined by schools as measured by the California English Language Development Test (California Department of Education, 2011a,b). Federal free/reduced price lunch program eligibility provides a crude measure of student socioeconomic status.

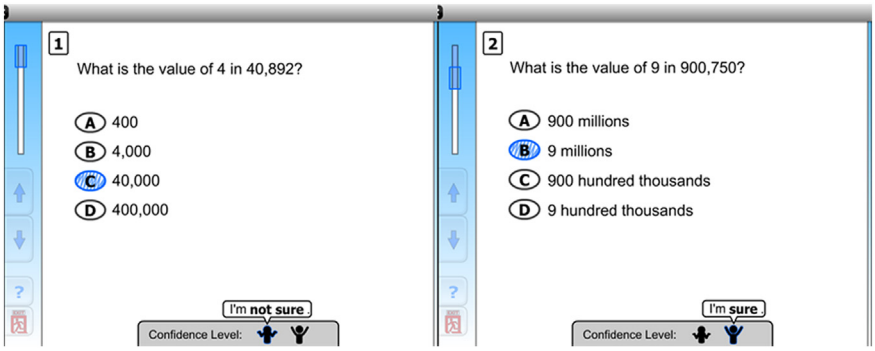

Fig. 3. Quiz questions appearing in ST Math. Students select their answer and then indicate confidence by selecting the appropriate icon. Here, the student on left is underconfident and on the right, is overconfident.

2.3. Analysis

2.3.1. Which measures of calibration can accommodate real-world data of accuracy and confidence judgments?

To answer this first research question, quiz question data were aggregated across all the objectives, pre and posttests, to provide the largest possible sample of questions. After aggregating, quadrants A through D were summed to represent each student's quadrant totals for each combination of confidence and accuracy. As a first step to understanding calculability of calibration measures, it was necessary to understand the presence of zero quadrants and whether quiz size and student characteristics were associated with these.

For this analysis, logistic regressions were calculated to predict the likelihood of a zero in each quadrant based on the number of questions answered. It was expected that student characteristics (e.g., gender, ethnicity) would be related to accuracy and confidence as well as number of questions answered (see Section 1.3). In this way, number of questions was confounded with student characteristics, both those measured and unmeasured. To address this issue, additional analyses were conducted using data from only those students who had completed a substantial portion of the curriculum (200 questions). For each student in this sample, 200 randomly drawn (without replacement) datasets of 25, 50, 75, 100, and 150 questions each were created, and from these, the percentage of students with zero quadrants was examined. Random selections of questions were chosen to also control for variation in question difficulty, hypothesized to be related to both student accuracy and confidence. This analysis examined the possibility of incalculable measures due to zero quadrants by using a range of realistic quiz lengths (see Nietfeld et al., 2006; Schraw et al., 2011).

After examining the possibility of zero quadrants, each of the ten measures described in Schraw et al. (2013) was analyzed to determine the proportion of students for whom each measure was not calculable. For this analysis, the full sample of student data with varying completion rates was utilized. For comparison, the measures were then recalculated after first adding 1 to each quadrant, approximating the procedure suggested in Hautus (1995), so that no students would have zero quadrants and all measures would be calculable. Means from this altered sample were compared to those from the unaltered sample.

2.3.2. Among calibration measures, which display the greatest predictive validity?

To answer this second research question, the data were first limited to students for whom each of the ten measures was calculable from pretest data only. Separate regressions were conducted to examine the association between pretest calibration and posttest accuracy for each measure, controlling for pretest accuracy, student grade level, number of quizzes completed, and student demographic variables. An additional model was examined considering Sensitivity and Specificity together, as recommended by Schraw et al. (2013). These analyses were replicated with the full sample of students with the measures adjusted to eliminate zero quadrants.

3. Results

3.1. Accommodation of real-world data

3.1.1. Zero quadrants

Within the ST Math curriculum, not all students took all quizzes—because of the self-paced nature of the program, students may not have reached the final objectives. The mean number of questions completed by students was 156.56 (SD = 73.43), differing slightly between the grade levels (Appendix Table 1). Fig. 4 displays the distribution of each quadrant within the 2 × 2 contingency table of accuracy and confidence. Any given student, however, may have a distribution of confidence and accuracy that filled only some of these quadrants. As noted above, due to the nature of calibration calculations, a zero in any quadrant may make measures of calibration incalculable. Of the 4281 students, 796 (19%) had zeroes in at least one quadrant: less than one percent of students had zeroes in quadrants A or B, indicating that most students had at least one question on which they were confident and correct and at least one question on which they were confident, but not correct. However, 15% of students had no unconfident and correct answers (quadrant C), and 9% had no unconfident and incorrect answers (quadrant D). Six percent of students had zeroes in two quadrants, less than one percent of students had zeroes in three quadrants.

A. Confident & Correct 56%	B. Confident & Incorrect 24%
C. Not Confident & Correct 8%	D. Not Confident & Incorrect 12%

Fig. 4. Distribution of combinations of confidence and accuracy within the actual ST Math quiz data. Compare with Schraw et al. (2013) simulated data where 62.5% of data was in cell A and 12.5% each in cells B through D.

Separate logistic regressions were estimated to predict the likelihood of having a zero in quadrants C or D based on number of questions completed. Regressions were not estimated for quadrants A and B because few students had zeros in these quadrants. Results are presented as odds ratios and marginal effects in Table 3. Students who completed more questions were less likely to have zeroes. For quadrant C, at the mean of total questions completed (157 questions), completing one more question was associated with a 0.10 percentage point decrease in the probability of having a zero quadrant. This indicates that a student would have to answer 150 more questions (307 total) to bring her probability from 15%[1] to nothing. Similarly, for quadrant D, at the mean of total questions completed, completing one more question was associated with a 0.10 percentage point decrease in the probability of having a zero quadrant. A student would have to answer another 90 (247 total) questions to bring her probability from 9% to nothing. This indicates that quizzes would have to be extremely impractical lengths of near 300 questions for information to be in all quadrants, rendering all calibration indices in Table 1 calculable.

As hypothesized, student factors were related to the number of quiz questions completed. Statistically significant associations emerged between these factors and question completion: in all grades but fifth grade, boys completed more questions than girls and those students eligible for free/reduced price lunch completed fewer questions than those who were not. In all grades, Asian students completed more questions than Hispanic students, and in all grades but third, ELLs completed fewer questions than non-ELLs.

To explore the association between number of questions completed and likelihood of having a zero quadrants without the

[1] 15% is the mean number of quadrant C zeroes using the entire dataset with a question range from 5 to 280, depending on grade—not necessarily the number of quadrant C zeroes at the mean question number, 157. Likewise, the mean number of quadrant D zeroes using the entire dataset is 9%.

Table 3
Odds ratios and marginal effects from logistic regression of zeroes in quadrants C and D by total number of questions.

N = 4281	C. Not Conf. & Correct		D. Not Conf. & Incorrect	
	Odds Ratio	Marg. Effects	Odds Ratio	Marg. Effects
Total No. Questions	0.994***	−0.001***	0.993***	−0.001***
	(0.001)	(0.0001)	(0.001)	(0.0001)
Grade 2	1.29*	0.033*	1.395*	0.028*
	(0.149)	(0.016)	(0.203)	(0.013)
Grade 3	1.015	0.002	1.002	0.0001
	(0.131)	(0.016)	(0.169)	(0.013)
Grade 5	1.055	0.007	1.470**	0.032**
	(0.119)	(0.014)	(0.199)	(0.012)
Constant	0.430***		0.227***	
	(0.047)		(0.031)	

Note. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Standard errors in parentheses. Grade 4 is the reference group.

confounding factor of student progress, the data were limited to those of students who had completed at least 200 questions. Demographic information on this reduced sample of 1341 students is provided in Appendix Table 2. This sample represented 31% of the original sample and varied across grades from 22% of fifth graders to 37% of second graders. After randomly drawing 200 datasets of 25, 50, 75, 100, and 150 questions from these data, the mean percentages of students who were missing values from each quadrant were examined, as were the 98% confidence intervals around these means. For each quadrant, the more questions that were used, the less likely there was a zero value in that quadrant. Using the mean number of zeros for each of the randomly drawn datasets, at 25 questions, less than 1% of students were without at least one question on which they were both correct and confident (quadrant A). Quadrant B, those questions which students got incorrect, but indicated confidence, contained a little more missing data. At 25 questions, between 3.69% and 8.17% of students, depending on grade, did not have any answers that fell in this quadrant. This number dropped to between 0.01% and 4.87% at 150 questions. Quadrants C and D, the quadrants representing student judgments of uncertainty, were missing from a large proportion of students. Approximately 40% of students never made judgments of uncertainty when they had the correct answer (quadrant C) in a randomly drawn sample of 25 questions. At 150 questions, this number dropped to between 9.03% and 16.59%, depending on grade. Approximately 30% of students never made judgments of uncertainty when they had the incorrect answer (quadrant D) at 25 questions. At 150 questions, this dropped to between 5.07% and 15.02%, depending on grade. Full results are available in Appendix Tables 4 through 7.

3.1.2. Measures of calibration

Table 4 presents the descriptive statistics from the calculation of the ten measures of calibration described in Section 1.2.1., Table 1, and taken from Schraw et al. (2013). These statistics are based on the full sample of 4281 students. As predicted from the presence of zero quadrants within these data, not all measures could be calculated for all students. Sensitivity, Specificity, Simple Match, G Index, and Sokal Reverse could be calculated for almost all the students—98% or more of the students in the sample had valid data for these measures. Odds Ratio, Gamma, and Phi suffered moderately from the presence of zero quadrants. For example, in fourth grade, which is the largest sample of students (N = 1522), Odds Ratio could only be calculated for 85% of the students and Gamma and Phi for 92% of the students each. Discrimination seemed to be most affected by zero quadrant scores: only 83% of fourth graders had valid Discrimination scores.

3.2. Predictive validity

To examine which of these measures had the most predictive validity, the data were separated by pre and posttest. The ten measures of calibration were recalculated using only the pretest measures, aggregated across all quizzes taken by the students. Limiting the data to those students who had at least one pretest reduced the sample by three students (N = 4278). The data were then limited to those students who had calculable values for each of the ten calibration measures, resulting in a new dataset of 3089 students, or 72% of those with pretest data. The analysis sample was further limited to the 3033 students who had complete demographic information. Student demographics and calibration measure descriptive statistics on this sample are presented in the Appendix.

As a first step, zero-order correlations were calculated to compare each of the ten measures, pretest and posttest accuracy, and pretest confidence (Appendix Table 9). With one exception (Kappa with Sensitivity), all measures of calibration were correlated to levels of statistical significance of $p < 0.05$, with correlations ranging from 0.07 (Sensitivity and Phi) to a perfect correlation between G Index and Simple Match. Sensitivity had low correlations (absolute values ranging from 0.01 to 0.23) with all measures other than Specificity, with which it had a strong inverse correlation of −0.73.

Table 5 displays the results from regressions of posttest accuracy on pretest accuracy and each measure of calibration considered separately. Full tables with results from control variables (gender, grade, ethnicity, language and free/reduced priced lunch status, and number of questions completed) are available in the Appendix. In

Table 4
Ten common measures of calibration calculated for entire sample from all available data.

N = 4281	2nd Grade			3rd Grade			4th Grade			5th Grade		
	Mean	SD	%Valid	Mean	SD	%Valid	Mean	SD	%Valid	Mean	SD	%Valid
Sensitivity	0.86	0.17	99.89%	0.87	0.16	100.00%	0.85	0.16	100.00%	0.87	0.15	100.00%
Specificity	0.31	0.27	99.34%	0.29	0.24	100.00%	0.36	0.27	99.67%	0.31	0.24	98.64%
Simple Match	0.66	0.12	100.00%	0.67	0.14	100.00%	0.65	0.13	100.00%	0.68	0.12	100.00%
Gamma	0.54	0.41	92.57%	0.54	0.41	96.06%	0.57	0.37	94.22%	0.53	0.42	91.38%
G Index	0.33	0.25	100.00%	0.33	0.27	100.00%	0.3	0.25	100.00%	0.35	0.25	100.00%
Odds Ratio	5.42	5.52	82.30%	5.84	6.62	86.95%	5.72	5.43	85.48%	5.32	6.07	82.46%
Kappa	0.17	0.16	99.45%	0.18	0.16	100.00%	0.21	0.17	99.67%	0.18	0.16	99.42%
Phi	0.22	0.16	92.57%	0.22	0.16	96.06%	0.25	0.16	94.22%	0.23	0.16	91.38%
Sokal Reverse	0.57	0.11	100.00%	0.57	0.12	100.00%	0.58	0.11	100.00%	0.56	0.12	100.00%
Discrimination	0.76	0.47	79.34%	0.76	0.48	83.67%	0.82	0.46	83.25%	0.77	0.43	78.59%
N	915			812			1522			1032		

Note. Includes all students in the sample combining questions in both pre and posttests. %Valid represents the percent of students for whom the given measures is calculable.

Table 5
Regression of posttest Accuracy (percentage of items correct) on pretest calibration and Accuracy for ten measures of calibration.

Diagnostic efficiency & agreement measures (Panel A)								
N = 3033		(1)	(2)	(3)	(4)		(5)	(6)
		Acc. Only	Sensitivity	Specificity	Sensitivity	Specificity	Simple Match	G Index
Measure(s)	B		0.046***	−0.003	0.098***	0.047***	0.074***	0.037***
	SE		(0.009)	(0.006)	(0.014)	(0.010)	(0.018)	(0.009)
	Beta		0.052***	−0.004	0.109***	0.074***	0.056***	0.056***
Pretest Acc.	B	0.818***	0.803***	0.818***	0.789***		0.779***	0.779***
	SE	(0.012)	(0.012)	(0.012)	(0.012)		(0.015)	(0.015)
	Beta	0.758***	0.744***	0.758***	0.731***		0.721***	0.721***
Constant	B	0.147***	0.121***	0.148***	0.066***		0.121***	0.158***
	SE	(0.009)	(0.011)	(0.010)	(0.015)		(0.011)	(0.009)
R2		0.697	0.699	0.697	0.702		0.698	0.698

Association, Binary Distance, and Discrimination (Panel B)							
N = 3033		(7)	(8)	(9)	(10)	(11)	(12)
		Gamma	Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
Measure(s)	B	0.028***	0.0004*	0.046***	0.049***	−0.081***	0.017***
	SE	(0.005)	(0.0002)	(0.010)	(0.010)	(0.022)	(0.003)
	Beta	0.057***	0.021*	0.049***	0.054***	−0.052***	0.055***
Pretest Acc.	B	0.798***	0.812***	0.804***	0.803***	0.781***	0.799***
	SE	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)	(0.012)
	Beta	0.739***	0.752***	0.745***	0.744***	0.723***	0.740***
Constant	B	0.144***	0.147***	0.143***	0.142***	0.215***	0.145***
	SE	(0.009)	(0.009)	(0.009)	(0.009)	(0.020)	(0.009)
R2		0.700	0.697	0.699	0.699	0.698	0.699

Note. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Unstandardized (B) and standardized (Beta) regression coefficients presented. Standard errors from unstandardized regressions are in parentheses. Control variables included number of pre and posttest questions completed, grade, gender, ethnicity, language and free/reduced priced lunch statuses. The reference group comprises students who were females in fourth grade, Hispanic, Non-ELL, and not on free lunch. Sample limited to those students who have non-missing values for each of the ten measures of calibration as described in Schraw et al. (2013).

Model 1, before the calibration measures were added, pretest accuracy and student demographics explained 69.7% of the variance in posttest accuracy. Adding an individual measure of calibration brought this, at most, to 70% of the variance as is seen in Model 7. Of the single-measure models, the Gamma model explained the most variance and also had the largest standardized regression coefficient, at 0.057. This indicates that a one standard deviation increase in Gamma was associated with less than one tenth of a standard deviation increase in aggregate posttest accuracy with pretest accuracy controlled. The combined Sensitivity/Specificity model produced a slightly larger R-squared than the Gamma model, explaining 70.2% of the variance ($\beta = 0.109$, Sensitivity, $\beta = 0.074$, Specificity).

Limiting the sample to only those students who had all ten measures calculable may have biased the dataset. As an alternative analysis, the data were modified to ensure that all students with at least one valid pretest would have data in all four quadrants before the ten measures were calculated. To do this, a 1 was added to each quadrant. Absolute differences between these new values and non-adjusted values were small (largely below 0.10, except for Odds Ratio), but in standard deviation units, ranged from less than 2/10ths of a standard deviation (e.g., Simple Match, G Index) to 4/10ths of a standard deviation (e.g., Odds Ratio). To determine whether these differences influenced the predictive validity of each measure, the regression of posttest score on pretest accuracy, calibration, and controls was conducted for these newly calculated measures.

Not all of the 4278 students with pretest data also had demographic data, and so, as in the prior analyses, the data were limited to those with non-missing data on the demographic covariates, resulting in a sample of 4144 (97% of the full pretest sample). Regression results were similar to those from the reduced sample and are presented in the Appendix. A model without any calibration measures explained 67% of the variance in posttest scores. As in the prior analyses, of the single calibration measures, Gamma added the most explained variance, adding an additional 0.3%. However, within these data, models with Kappa and Phi also added an additional 0.3%. Unlike the limited sample models, in these regressions, G Index emerged as the strongest single predictor ($\beta = 0.061$), although differences in magnitude of standardized regression coefficients were small between many of the measures: six of the measures had betas within 0.01 of each other. Replicating the prior analysis, the combined model with Sensitivity and Specificity explained more variance than a single-measure model (67.5%). Considered in a model together, Sensitivity ($\beta = 0.114$) and Specificity ($\beta = 0.094$) had stronger associations with posttest score than did any other individual measure of calibration.

4. Discussion

This study set out to answer two research questions: (1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments? and (2) Among these measures, which display the greatest predictive validity? These questions were answered with data rarely used in comparisons of multiple measures of calibration: data from authentic learning tasks with a large number of questions and large sample size.

4.1. Accommodation of real-world data

4.1.1. Zero quadrants

Even in the preliminary analyses, differences were apparent between these data and the type of simulated data often used for measurement comparison studies. The students taking the quizzes within ST Math did not have accuracy and confidence judgments that were evenly distributed among cells B through D as most simulated data studies assume. The majority of responses were in cell A (56%) which did somewhat replicate studies meant to

approximate realistic conditions (Nietfeld et al., 2006; Schraw et al., 2013). However, cell C (not confident and correct) appeared the least often (8%), indicating that few of the student responses displayed underconfident patterns. Prior research indicated concern that cell D (not confident and incorrect) would be the option most likely to remain unchosen by participants (Schraw et al., 2011), but in the current comparison, cell C was the most likely cell to be left empty.

A zero in at least one quadrant affected 19% of the students using the largest possible sample of questions and participants (approximately 156 questions for each student). Logistic regression results indicated that for this sample, tests of over 300 questions would likely be needed to avoid zero quadrants and the resulting measure distortion. This is a far cry from the 25 questions suggested by Schraw et al. (2011). Schraw and colleagues based this suggestion on data simulated to replicate moderate difficulty (75% accuracy), and indicated that the more difficult the test, the more equal the distribution among the quadrants and the less likely zero quadrants would be. Based on the accuracy of the current sample, the ST Math quizzes appeared *more* difficult than Schraw's simulated data (64% accuracy), leaving a larger number of responses available for distribution in quadrants B through D. However, the majority of the 44% of responses not within quadrant A were in quadrant B (confident and incorrect), indicating strong overconfidence among the student participants. This overconfidence is typical in young students (Pajares & Kranzler, 1995; Pressley, Levin, Ghatala, & Ahmad, 1987). As Schraw and colleague's recommendation was based on simulated data intended to approximate adult behavior, it may not be applicable to measures of calibration in children.

It could be that the age of the children may not be the only thing causing these disparate results. Completion of questions was related to a number of demographic student characteristics. It is also possible that completion could have been related to characteristics such as math proficiency or familiarity with the format within ST Math—things that also affect the students' ability to marshal metacognitive resources and make accurate confidence judgments (Alexander & Murphy, 1999; Kruger & Dunning, 1999). Additionally, because of the structure of ST Math, the full sample included more questions from the start of the curriculum, reducing the external validity of the findings. As a step toward removing this confound and increasing external validity, sample tests of varying question lengths were created through random selection of responses from among those students who had completed at least 200 questions. By limiting the data to only those students who had completed 200 questions, examinations could be made within a group of students more likely to be similar to each other, and by drawing the random datasets, the difficulty of the questions was more randomly distributed. However, it should be noted that this sample was demographically different from the original sample—exact differences are available in Appendix Table 2. This permitted analysis of small-sized quizzes (e.g., 25 questions) without having to rely on questions from a small sample of objectives that may have been easier or harder than the other objectives. Examination of these data suggested similar patterns to those observed in the data overall: quadrant C appeared the most problematic, with 40% of students missing data from this quadrant in 25-question quizzes. At 150 questions, quadrant C remained the most problematic: averaged across the grades, 15% of the students had a zero in quadrant C and 10% had a zero in quadrant D. Despite the suggestion from the logistic regression results indicating that zero quadrants would be eliminated at around 300 questions, it cannot be said for certain that with this population and subject-matter, zero quadrants would be eliminated even at 1000 questions, the number typically used in simulation experiments to approximate a test assumed to be problem-free.

4.1.2. Measure calculation

Even at test lengths of 150 questions, zeroes in quadrants were likely to be a problem. However, not all zeroes would result in undefined measures. For example, Gamma could be calculated with zeroes in quadrant C or D, as long as both were not missing.[2] Using all the available data, all ten measures could be calculated for most students. However, both Gamma and Discrimination suffered from undefined values. For Gamma, between 4 and 9% of the cases, depending on grade, were undefined or otherwise incalculable (compare with only 0.1% in a 20-item test of similar difficulty in Schraw et al., 2011). As in Kuch (2012), there were more undefined values for Discrimination than for Gamma and the other measures. For second and fifth graders, over 20% of the values for Discrimination were undefined. This is close to the 17% Kuch (2012) found for 20-item tests, but the majority of the test-takers in the current study took well over 20 questions: fewer than 3% of the students in these data completed fewer than 20 items; the majority completed more than 150. These issues further highlight the difficulty in translating measurement and calculation decisions from simulated studies to those with authentic data and, to some extent, across contexts within non-simulated studies. Although the current paper addressed these issues with respect to calibration, there is reason to believe differences between simulated and authentic data may also be present in studies of other topics. With Monte Carlo and other data simulation studies on the rise (Mishra, Koehler, & Greenhow, 2015), educational psychologists may wish to be mindful of the example of mismatch illustrated herein.

4.2. Predictive validity

To our knowledge, this is the first study to compare the relative predictive validity of these ten commonly-used measures of calibration using authentic education data. Some differences between the measures in these data and prior simulated data should be highlighted. The ten measures are assumed to be correlated, except for Sensitivity and Specificity, which are assumed to be orthogonal and have been shown to be so in simulated data (Schraw et al., 2013, but cf. Gutierrez, Schraw, Kuch, & Richmond, 2016 showing an inverse correlation in non-simulated data). However, within these data, Sensitivity and Specificity were inversely and statistically significantly correlated and this correlation was relatively strong. Sensitivity and Specificity were more highly correlated with each other than with any of the other measures—although Specificity had moderate to strong correlations with Kappa, Phi, and Discrimination. Other researchers have advocated for Gamma as the gold standard measure of calibration, partly because of its assumed correlation with other measures (Nelson & Narens, 1990; but cf; Schraw et al., 2013). Gamma did have strong correlations with all measures except for Sensitivity and Specificity, but it was not alone—many of the measures were as highly intercorrelated.

In considering the relative predictive strength of the measures, Gamma was the strongest single predictor, in line with the Nelson and Narens (1990) model suggesting a single monitoring process, but the predictive advantage was very small. The model combining Sensitivity and Specificity explained the most variance in post-test scores and had the highest beta values. This is in agreement with Schraw et al. (2013) suggestion that Sensitivity and Specificity represent unique aspects of monitoring (see also Feuerman & Miller, 2008). Within these data, it appears that a model that accounts separately for students' knowledge of what they do know (Sensitivity) and what they do not know (Specificity) is more

[2] The ability to calculate the measure does not preclude distortion of the measure due to one zero quadrant (see Kuch, 2012).

powerful than one that includes a measure that conflates the two. Even still, the combined model only explained three tenths of a percent more variance than the models with the next highest R-squared values.

Given the high degree of association between each of the measures, their similar levels of predictive validity may not be surprising. What may be surprising is the small amount of variance in posttest accuracy explained by calibration measures. Zero-order correlations between calibration and achievement were in line with prior research—most with an absolute value about or above 0.30 (e.g., Barnett & Hixon, 1997; Desoete & Roeyers, 2006). However, in much of this prior work, the same test is used to measures calibration and achievement (e.g., Bol et al., 2010), or correlations between pretest calibration and posttest performance are examined without controlling for pretest performance (e.g., Barnett & Hixon, 1997). If the accuracy of metacognitive judgments is indicative of a regulatory process not entirely subsumed by prior achievement, it should uniquely contribute to future performance net of prior performance. There *was* unique variance in posttest achievement explained by pretest calibration, but although statistically significant, beta values were mostly under 0.10. Nevertheless, these small values may have cumulative impact in student learning: students who are able to monitor accurately should be able to adjust their strategies and resources for greater success (Nelson, 1996; Pintrich, 2004; Winne, 2001). If this student monitoring accuracy is a dispositional characteristic, some of this association may have been controlled away with our inclusion of a prior achievement measure.

4.2.1. Replication of results by correcting for zero quadrants

In a replication of the regression analyses, a 1 was added to each quadrant to ensure that all measures were calculable and that zero quadrants did not otherwise distort the values of the calibration measures. Following the procedure suggested in Hautus (1995), a value was added to each quadrant instead of only to the missing quadrants. Differences in the means of measures between this sample altered for non-missing data and the sample limited to only non-missing participants were not negligible (close to 4/10ths of a standard deviation for Odds Ratio, Gamma, and Discrimination). These differences did not translate to large differences in predictive validity between measures, however. Betas and R-squared values between the models were close, and, agreeing with the limited sample analysis, the model including both Sensitivity and Specificity explained the most variance and had the largest standardized regression coefficients.

4.3. Limitations

The greatest strength to this study, that the data were taken from an authentic learning task completed by real students, is also a limitation. Because the data were real and suffered from real-world problems, they do not exactly match large simulation studies. Had a test of 1000 questions been administered to the students, it may have been possible to make comparisons similar to those conducted in simulated studies, but such a test is impractical in a single administration. If it were administered in smaller chunks over the course of a year, it would be likely to contain the same types of missingness found in the data herein. Although the quizzes administered within ST Math allowed examination of data of a type and scope not studied previously within the calibration literature, the results may be limited to similar populations and materials. There are reasons to believe there are domain and age differences in calibration (Bong, 1999; Jonsson & Allwood, 2003), and these differences may extend to the calculability and predictive validity of the different measures.

5. Conclusion

Prior recommendations regarding the measurement of calibration are based largely on simulated data. Data in the current study, taken from student interactions with authentic mathematics learning tasks, do not behave as simulated data do: distribution among the four quadrants is not even, and patterns of missingness do not mirror those found in simulated studies. These differences have real implications for the calculability of many of the measures commonly used in calibration research. Researchers may wish to avoid measures like Gamma or Discrimination and instead rely upon measures more robust to missing quadrants, such as G Index. Outside of such practical considerations, measure selection can also be guided by predictive validity. Results of this study supported assertions by Schraw et al. (2013) that Sensitivity and Specificity, when used together, best represent metacognitive accuracy and will likely be the most powerful predictors of achievement in calibration studies. These two measures have a long history of use in clinical research, but until recently had not been used in the measurement of calibration within educational settings. Although the results herein may recommend their use, more work is needed to understand the actual processes underlying determinations of confidence and uncertainty, especially in light of the high correlation between the measures within these data.

Acknowledgments

Funding: This work was supported by the Institute for Education Sciences [grant number R305A090527]; and the National Science Foundation [grant number DGE-0808392]. I would like to thank George Farkas, Greg Duncan, Jacquelynne Eccles, Elizabeth Loftus, Deborah Vandell, AnneMarie Conley, and John Nietfeld, who provided feedback on this study and manuscript. I would also like to thank the participating students, teachers, and schools.

Appendix A. Supplementary data

Supplementary data related to this chapter can be found at http://dx.doi.org/10.1016/j.learninstruc.2016.10.006.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3. http://dx.doi.org/10.1016/j.learninstruc.2012.10.003.

Alexander, P. A., & Murphy, P. K. (1999). Nurturing the seeds of transfer: A domain-specific perspective. *International Journal of Educational Research, 31*(7), 561–576. http://dx.doi.org/10.1016/S0883-0355(99)00024-5.

Allwood, C. M., Jonsson, A.-C., & Granhag, P. A. (2005). The effects of source and type of feedback on child witnesses' metamemory accuracy. *Applied Cognitive Psychology, 19*(3), 331–344. http://dx.doi.org/10.1002/acp.1071.

Bandura, Albert (1986). *Social foundations of thought and action : A social cognitive theory.* Englewood Cliffs N.J: Prentice-Hall.

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90*(3), 170–174. http://dx.doi.org/10.2307/27542087.

Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction, 20*(5), 372–382. http://dx.doi.org/10.1016/j.learninstruc.2009.03.002.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133–151. http://dx.doi.org/10.2307/20152656.

Bol, L., Riggs, R., Hacker, D. J., Dickerson, D. L., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*, 81–96.

Bong, M. (1999). Personal factors affecting the generality of academic self-efficacy judgments: Gender, ethnicity, and relative expertise. *The Journal of Experimental Education, 67*(4), 315–331. http://dx.doi.org/10.1080/00220979909598486.

California Department of Education. (2011a). *California english language development test [test website].* Retrieved from http://www.cde.ca.gov/ta/tg/el/.

California Department of Education. (2011b). *California STAR report*. Retrieved from http://star.cde.ca.gov/star2011.

Chen, P. (2002). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77–90.

Desoete, A., & Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. *Learning and Instruction, 16*(1), 12–25.

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: Sensitivity, Specificity and kappa. *Journal of Evaluation in Clinical Practice, 14*(5), 930–933. http://dx.doi.org/10.1111/j.1365-2753.2008.00984.x.

Gutierrez, A. P., & Price, A. F. (2016). Calibration between undergraduate students' prediction of and actual performance: The role of gender and performance attributions. *The Journal of Experimental Education*, 1–15. http://dx.doi.org/10.1080/00220973.2016.1180278.

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1–10.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51. http://dx.doi.org/10.3758/BF03203619.

Huff, J., & Nietfield, J. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*(2), 161–176.

Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences, 34*(4), 559–574. http://dx.doi.org/10.1016/S0191-8869(02)00028-4.

Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the performance of business students on examinations. *Journal of Education for Business, 79*(4), 217–224. http://dx.doi.org/10.3200/JOEB.79.4.217-224.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kuch, F. (2012). *A comparison of bias in four measures of monitoring accuracy* (Doctoral dissertation). Retrieved from http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=2586&context=thesesdissertations.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development, 82*(6), 1778–1787. http://dx.doi.org/10.1111/j.1467-8624.2011.01649.x.

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of Sensitivity and response bias. *Psychonomic Bulletin & Review, 3*(2), 164–170. http://dx.doi.org/10.3758/BF03212415.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(2), 509–527. http://dx.doi.org/10.1037/a0014876.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition, 38*, 441–451. http://dx.doi.org/10.3758/MC.38.4.441.

Miller, J. (1996). The sampling distribution of d′. *Perception & Psychophysics, 58*(1), 65–72. http://dx.doi.org/10.3758/BF03205476.

Mishra, P., Koehler, M. J., & Greenhow, C. (2015). The work of educational psychologists in a digitally networked world. In L. Corno, & E. M. Anderman (Eds.), *Handbook of educational psychology* (pp. 29–40). New York: Taylor & Francis.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*(2), 102–116. http://dx.doi.org/10.1037/0003-066X.51.2.102.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742108600535.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*(2), 258–271. http://dx.doi.org/10.1177/0013164404273945.

Ots, A. (2013). Third graders' performance predictions: Calibration deflections and academic success. *European Journal of Psychology of Education, 28*(2), 223–237. http://dx.doi.org/10.1007/s10212-012-0111-z.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 20*(4), 426–443. http://dx.doi.org/10.1006/ceps.1995.1029.

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The Journal of Experimental Education, 65*(3), 213–228. http://dx.doi.org/10.1080/00220973.1997.9943455.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Elsevier.

Pintrich, P. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–407. http://dx.doi.org/10.1007/s10648-004-0006-x.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*(1), 96–111. http://dx.doi.org/10.1016/0022-0965(87)90053-1.

Rutherford, T., Farkas, G., Duncan, G., Burchinal, M., Graham, J., Kibrick, M., … Martinez, M. E. (2014). A randomized trial of an elementary school mathematics software intervention: Spatial-Temporal (ST) Math. *Journal of Research on Educational Effectiveness, 7*(4), 358–383. http://dx.doi.org/10.1080/19345747.2013.856978.

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

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

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.

Schraw, G., Kuch, F., Gutierrez, A. P., & Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology, 106*(4), 1192. http://dx.doi.org/10.1037/a0036653.

Schraw, G., Kuch, F., & Roberts, R. (2011, April). *Bias in the gamma coefficient: A Monte Carlo study*. Calibrating Calibration: Conceptualization, measurement, calculation, and context. Symposium conducted at the Annual meeting of the American Educational Research Association, New Orleans, LA.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. http://dx.doi.org/10.3102/00346543075003417.

Stone, N. (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.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. http://dx.doi.org/10.1037/0022-0663.95.1.66.

Tobias, S., & Everson, H. T. (1998, April). *Research on the assessment of metacognitive knowledge monitoring*. Paper presented at the annual convention of the American Educational Research Association, San Diego.

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the child behavior checklist and strengths and difficulties questionnaire: A systematic review. *Child and Adolescent Mental Health, 13*(3), 140–147. http://dx.doi.org/10.1111/j.1475-3588.2007.00461.x.

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*(4), 173–187. http://dx.doi.org/10.1207/s15326985ep3004_2.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 279–306). Hillsdale, NJ: Erlbaum.

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.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. http://dx.doi.org/10.1037/0022-0663.81.3.329.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183.