The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Mathematics Self-Efficacy and Mathematical Problem Solving: Implications of Using Different Forms of Assessment

Frank Pajares & M. David Miller

To cite this article: Frank Pajares & M. David Miller (1997) Mathematics Self-Efficacy and Mathematical Problem Solving: Implications of Using Different Forms of Assessment, The Journal of Experimental Education, 65:3, 213-228, DOI: 10.1080/00220973.1997.9943455

To link to this article: https://doi.org/10.1080/00220973.1997.9943455

Published online: 14 Jul 2010.

The Journal of Experimental Education, 1997, *65*(3), 213–228

Mathematics Self-Efficacy and Mathematical Problem Solving: Implications of Using Different Forms of Assessment

FRANK PAJARES
Emory University

M. DAVID MILLER
University of Florida

ABSTRACT. The mathematics self-efficacy and problem-solving performance of 327 middle-school students were assessed using two forms of assessment (traditional multiple-choice vs. open-ended fill-in-the-blank). The purpose was to determine whether varying the assessment format would influence students' self-efficacy judgments or alter the relationship between self-efficacy and performance. No differences in self-efficacy resulted from the different forms of assessment. Students who took the multiple-choice performance test obtained higher scores than did students who took the open-ended test. The latter group had poorer calibration, that is, the degree to which students' judgments of their capability reflect their actual competence. This finding suggests that students' self-perceptions of their mathematics capability may be less accurate than has previously been reported or that students' familiarity with traditional assessment formats creates an expectancy of a performance task that is multiple choice in nature; this expectancy influences self-efficacy judgments regardless of the format used to assess confidence. Differences in the format for assessing self-efficacy and performance altered the predictive utility of self-efficacy judgments. These differences must be accounted for in subsequent studies before researchers can make sound generalizations about the strength of the self-efficacy/performance relationship or the accuracy of students' self-perceptions.

SOCIAL COGNITIVE THEORISTS have hypothesized that students' self-efficacy beliefs, that is, their judgments of their capability to accomplish specific academic tasks, are important determinants of academic motivation, choices, and performance (e.g., Bandura, 1986, 1997; Pajares, 1996b; Schunk, 1991). These beliefs of self-capability affect motivation by influencing the effort and persistence with which students engage in academic tasks, as well as the anxiety they experience. Social cognitive theorists also hypothesize that these beliefs of self-capability mediate the influence of other determinants of academic outcomes. In part, this is because the confidence that students have in their own capability helps determine what they do with the knowledge and skills they possess. Thus, social cognitive theorists maintain that the academic performance of students is determined in large measure by the confidence with which they approach academic tasks (Bandura, 1997; Schunk, 1991). For example, when students take a mathematics exam, the self-confidence that they experience as they read and analyze specific problems in part determines the amount of time and effort they put into solving those problems. Student with greater confidence work harder and longer and are less anxious. As a result, their chances of successful academic performance are enhanced.

Various researchers (e.g., Hackett, 1985; Pajares, 1996a; Pajares & Miller, 1994) have reported that students' judgments of their capability to solve mathematics problems are predictive of their actual capability to solve those problems. These judgments also mediate the influence of other predictors such as math background, math anxiety, perceived usefulness of mathematics, prior achievement, and gender. Math self-efficacy also has been shown to be as strong a predictor of mathematical problem-solving capability as general mental ability (Pajares & Kranzler, 1995), a variable generally found to be a powerful predictor of academic performance (Thorndike, 1986). Other researchers (Collins, 1982; Schunk, 1989, 1991) have reported that, when students approach academic tasks, those with higher self-efficacy work harder and for longer periods of time than do those with lower self-efficacy.

Researchers have usually assessed mathematics self-efficacy by asking students to use a Likert-type scale to indicate the strength of their confidence in solving various mathematics problems. Specifically, students are presented with a number of problems and asked to rate how confident they are that they can successfully solve each one. Efficacy instruments such as the Mathematics Self-Efficacy Scale (MSES; Betz & Hackett, 1983), for example, ask students to provide judgments of their capability to solve specific algebra or geometry problems correctly. However, whereas the subsequent performance assessment presents students with these problems in a traditional multiple-choice format, the efficacy instrument presents the problems without the multiple-choice options. Pajares and Miller (1995) have shown the importance of having a close link between efficacy judgments and the criterion task. Therefore, it is important to determine whether the form of self-efficacy assessment (multiple-choice or open-ended question format) differentially predicts performance and whether using multiple-choice or open-ended self-efficacy and performance assessments yields different relationships between self-efficacy and performance. In the present study, students were administered a self-efficacy measure in which math problems were presented either with possible multiple-choice answers or without answers in an open-ended format; performance assessments differed similarly. The resulting confidence/outcome correspondences among the four groups created by these distinctions (see Figure 1) should help clarify the relationship between self-efficacy and academic performances. Note that Group 1 represents the manner in which self-efficacy and performance are typically assessed.

Clarification of these issues is important for several reasons. First, it would inform the current controversy regarding alternative[1] and traditional assessment methods (Worthen, 1993) and the role that judgments of self-efficacy play in students' performance on them. For example, if simply presenting mathematics problems in one type of format produces greater confidence than presenting them in another does, it seems reasonable that students are better served by formats that increase self-efficacy beliefs, for these beliefs play a role in triggering stronger effort and persistence and lowering anxiety. Assuming that different forms of assessment can be developed to measure competence, the strength and level of self-efficacy that a particular form engenders should be one factor that is considered when effective assessment tests and programs are planned.

Second, study of the relationships shown in Figure 1 will have implications for

FIGURE 1. Groups created by varying efficacy and performance assessments.

		Self-efficacy assessment	
		OE	**MC**
Performance	**MC**	**OE self-efficacy** **MC performance** **(Group 1)** **Open-ended self-efficacy instrument; multiple-choice performance test.**	**MC self-efficacy** **MC performance** **(Group 2)** **Both instruments presented with multiple-choice answer format.**
	OE	**OE self-efficacy** **OE performance** **(Group 3)** **Both instruments presented with open-ended answer format.**	**MC self-efficacy** **OE performance** **(Group 4)** **Multiple-choice self-efficacy instrument; open-ended performance test.**

[1]We are aware that alternative assessment methods involve more than simply administering an open-ended test. However, mathematics tests with open-ended answers are one alternative to traditional assessment formats, and therefore we take the liberty of using the term *alternative* to refer to them.

the implementation of statewide assessments that are in new formats. For example, the state of Florida is currently developing new forms of assessment that are to be adopted statewide to match the objectives of *Blueprint 2000,* which map out the state's new educational mission. The hope is that new assessments will have a strong impact on both the self-efficacy and performance of students in the years to come. Specifically, it is hoped that the new assessments will "more closely mirror the teaching and learning process, resulting in greater instructional fidelity" (Miller & Seraphine, 1993, p. 119). Wiggins (1989) argued that if tests more closely match the instruction received, confidence will play a greater role in student learning and assessment. Thus, part of the rationale for alternative forms of assessment lies in their ability to increase students' confidence in their capability to do well on them.

A third benefit to be gained from research on formats for self-efficacy judgments and performance assessment involves the issue of calibration, that is, the degree to which students' judgments of their capability reflect their actual competence. Researchers (Hackett, 1985; Hackett & Betz, 1989; Pajares & Miller, 1994) consistently have found that most students are overconfident about their capability to solve mathematics problems. Pajares and Kranzler (1995) found that the vast majority of the high school students they tested demonstrated strong confidence in their ability to solve mathematics problems, but this confidence was not matched by their subsequent performance. Bandura (1986) argued that the most functional efficacy judgments are those that slightly exceed what one can actually accomplish. Excessive overconfidence, however, can result in "serious, irreparable harm" (p. 394).

It merits considering what role formats for self-efficacy judgments and performance assessment play in students' reported levels of confidence and in the observed correspondence between that confidence and their performance. It may be that the levels of overconfidence reported by researchers are at least partly a function of the formats that they used to measure self-efficacy and performance. Pajares and Kranzler (1995) recommended that research on the nature of the relationship between efficacy judgments and calibration is needed. They suggested that it may be more important to develop instructional techniques and intervention strategies to improve students' calibration than to attempt to raise their already overconfident beliefs. Improved calibration should result in better understanding by students of what they know and do not know so that they more effectively deploy appropriate cognitive strategies during the problem-solving process. If improved calibration is in part a function of self-efficacy assessment, then the assessment itself becomes a useful intervention to help students with this metacognitive capability.

Our purpose in the present study was to determine whether varying the form of assessment would influence students' self-efficacy judgments and affect the relationship between self-efficacy judgments and mathematics performance.

Specifically, our research objectives were (a) to determine how a multiple-choice versus fill-in-the-blank presentation of mathematics problems influences students' mathematics self-efficacy judgments and (b) to determine the relationship between these self-efficacy judgments and subsequent performance on traditional and alternative assessments of mathematical problem solving.

We also were interested in exploring the role that gender and math background play in the relationship between these variables. Researchers (Fennema & Sherman, 1978) have found that boys and girls do not differ in math performance during elementary school but that differences appear in middle school and increase throughout high school. When affective variables (e.g., self-confidence) are controlled, however, these gender differences lessen or disappear. This finding has led researchers to suspect that these affective variables are a key source of gender differences in mathematics achievement. When differences in math background (e.g., previous mathematics courses taken) are also controlled, even smaller gender differences in math achievement are found (Lapan, Boggs, & Morrill, 1989).

The relationship between gender and math self-efficacy has not been explored as thoroughly as that between gender and math performance. The authors of initial studies (Fennema & Sherman, 1978) suggested that boys were more confident in their math skills than were girls. More recent findings (Pintrich & De Groot, 1990) show that boys and girls have equal confidence during elementary school but that by middle school boys are more confident. Also, self-efficacy and form of assessment are known to vary by gender and math background (Miller & Legg, 1993). Researchers have reported that calibration is related to mathematics performance and general mental ability (Pajares & Kranzler, 1995), mathematics grade point average (Pajares, 1996a), and giftedness (Pajares, 1996a). Therefore, we examined gender differences as well as differences in mathematics background (algebra vs. prealgebra) to determine whether certain types of students express lower self-efficacy on particular forms of assessment and whether they exhibit a different relationship between self-efficacy and performance.

Method

Participants and Procedures

Participants consisted of 327 eighth-grade students enrolled in algebra (n = 199) and prealgebra (n = 128) classes in eight middle schools in a southern state. There were 178 girls and 149 boys. Instruments were administered by group in individual mathematics classes during two class periods. Within each class, the students were randomly assigned to one of four testing groups. During the first administration, students in Groups 1 and 3 were administered the version of the self-efficacy instrument in which answers to the individual math problems were left open-ended, whereas Groups 2 and 4 were administered a similar version with answers to the problems available in a multiple-choice format. Before completing the self-efficacy measure, the students were informed that they subsequently would be asked to solve the problems on which their confidence was being assessed, but they were strictly instructed to provide only confidence judgments at this time and not to attempt to solve the problems. On the second day, Groups 1 and 2 were administered the multiple-choice performance instrument, and Groups 3 and 4 received the same instrument with an open-ended format. The four groups correspond to those shown in Figure 1. Hereinafter the four groups are labeled as follows: OE Eff/MC Perf (Group 1), MC Eff/MC Perf (Group 2), OE Eff/OE Perf (Group 3), and MC Eff/OE Perf (Group 4).

Instruments

Math self-efficacy. The self-efficacy instruments asked students to express their level of confidence in successfully solving each of 30 mathematics problems drawn from the eighth-grade test of the Second Study of Mathematics (International Association for the Evaluation of Educational Achievement, 1985). The following are two sample items: "What is the square root of 12×75?"; "A shopkeeper has x pounds of tea in stock. He sells 15 pounds and then receives a new lot weighing $2y$ pounds. What weight of tea does he now have?" As we noted above, the OE Eff/MC Perf and OE Eff/OE Perf groups were presented the self-efficacy instrument with these problems in an open-ended format without answer options provided. Directions on the self-efficacy instruments asked students, "How confident are you that you could answer the following questions correctly without using a calculator?" The MC Eff/MC Perf and MC Eff/OE Perf groups were presented with the identical problems, but were provided with answer options in a 5-item multiple-choice format. The directions asked students, "How confident are you that you could answer the following multiple-choice questions correctly without using a calculator?"

In both versions of the self-efficacy instrument, students used a 6-point Likert-type scale that ranged from 1 *(no confidence at all)* to 6 *(complete confidence)* to rate the strength of their confidence that they could successfully solve each problem. Each student provided confidence judgments on each of the 30 mathematics problems. Final self-efficacy scores were the sum of confidence ratings for the 30 problems and ranged from a low of 30 to a high of 180. Both forms of the self-efficacy instrument were highly reliable: The multiple-choice and open-ended versions of the instrument each produced a Cronbach's alpha coefficient of .92.

Math performance. The performance instrument consisted of the same mathematics problems on which students provided their confidence judgments. The

OE Eff/OE Perf and MC Eff/OE Perf groups received a version of the test in a fill-in-the-blank format and were asked to write the answer in the blank that corresponded to each problem. The OE Eff/MC Perf and MC Eff/MC Perf groups were provided with 5-item multiple-choice answers and were asked to circle the correct one. Cronbach's alpha coefficient for the open-ended test was .87; the alpha coefficient for the multiple-choice test was .82.

Calibration. We used three measures of calibration. The first was the *mean bias* score described by Keren (1991), Schraw (1995), and Yates (1990). Bias reveals the direction of the errors in judgment and is computed by subtracting actual performance from predicted confidence. Performance was scored so that a correct answer was scored as 6 and an incorrect answer was scored as 1. Scores for the Likert-type scale for the self-efficacy instrument also ranged from 1 to 6. Thus, expressing no confidence (1) and answering incorrectly (1) would reflect zero bias (1 – 1), whereas the same lack of confidence with a correct answer would receive a bias score of –5 (1 – 6), indicating underconfidence. With this procedure, then, bias scores could range from –5 to +5. Scores larger than zero corresponded to overconfidence; scores less than zero corresponded to underconfidence. The 30 bias scores (1 for each item) were averaged to yield a mean bias score.

The second measure of calibration was *mean accuracy,* which was computed by subtracting the absolute value of each bias score from 5. This score reveals the magnitude of the judgment error, which could range from 0 (complete inaccuracy) to 5 (complete accuracy). The 30 accuracy scores (1 for each item) were averaged to yield a mean accuracy score. Similar procedures and their rationale are described elsewhere (Schraw, 1995; Schraw & DeBacker Roedel, 1994; Schraw, Dunkle, Bendixen, & DeBacker Roedel, 1995; Schraw, Potenza, & Nebelsick-Gullet, 1993).

The third measure of calibration was *item accuracy,* which is a score that reflects the number of items on which a student's confidence judgment and performance attainment concurred. Specifically, the score is the number of items on which the student expressed confidence (by marking 4, 5, or 6) and answered correctly plus the number of items on which the student expressed lack of confidence (by marking 1, 2, or 3) and answered incorrectly. This last measure of calibration was computed so as to permit comparisons with prior findings when this technique was used.

Results

Descriptive statistics for each treatment group's scores on the dependent variables are shown in Table 1. A multivariate analysis of variance (MANOVA) revealed a significant between-groups effect for the dependent variables of math-

TABLE 1
Descriptive Statistics for Variables, by Treatment Group

	Group							
	OE Eff/MC Perf		MC Eff/MC Perf		OE Eff/OE Perf		MC Eff/OE Perf	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Performance	18.78_a	5.20	19.36_a	5.14	14.78_b	6.09	15.70_b	5.82
Self-efficacy	146.31	18.52	144.79	20.38	141.61	22.55	145.91	18.40
Bias	0.75_a	0.84	0.60_a	0.78	1.26_b	0.86	1.25_b	0.84
Mean accuracy	3.17	0.60	3.28_a	0.62	2.98_b	0.58	3.04_b	0.62
Item accuracy	19.54	4.23	20.26_a	4.69	18.10_b	4.44	18.41_b	4.62

Note. Group means for a dependent variable (row) that are subscripted by different letters are statistically different (experimentwise $\propto \leq .05$) according to a Tukey HSD computed on an effect identified by MANOVA. OE = open-ended; MC = multiple-choice; Eff = self-efficacy; Perf = performance.

ematics performance, self-efficacy, mean bias, mean accuracy, and item accuracy, Wilks's lambda = .77, $F(12, 842) = 7.19$; $p < .0001$. Univariate analysis of variance (ANOVA) results showed group effects for all dependent variables with the exception of self-efficacy—students reported similar levels of confidence whether problems were presented to them in a multiple-choice or open-ended format, $F(5, 321) = 2.07$, $p > .07$. There were significant group effects for performance, $F(5, 321) = 41.60$, $p < .0001$; for mean bias, $F(5, 321) = 38.33$, $p < .0001$; for mean accuracy, $F(5, 321) = 32.01$, $p < .0001$; and for item accuracy, $F(5, 321) = 33.70$, $p < .0001$. Tukey's HSD test showed that the two groups that took the multiple-choice performance test significantly outperformed the groups that took the open-ended test. As a result of their lower performance scores, the groups that took the open-ended test also had higher mean bias scores, reflecting greater overconfidence. Differences in mean accuracy and item accuracy were found only between the group that took both measures in a multiple-choice format and the two open-ended performance groups.

Prediction of Mathematics Performance

Table 2 contains Pearson product-moment correlations between self-efficacy and performance scores for the full sample, for each treatment, for each algebra group, and for each gender group. We observed significant correlations across all treatment groups and for the algebra and prealgebra groups. To better understand the nature of the self-efficacy/performance relationship and of the gender differences in the correlations, we conducted multiple regression analyses to determine whether students' self-efficacy judgments across the four groups were predictive of their performance scores. The analyses controlled for the influence of enrollment level (algebra vs. prealgebra), gender, self-efficacy test format (multiple-

TABLE 2
Correlations Between Self-Efficacy and Problem-Solving Performance by Treatment Group, Level, and Gender

		Group			
Group	Total	OE Eff/ MC Perf	MC Eff/ MC Perf	OE Eff/ OE Perf	MC Eff/ OE Perf
Full sample	.49***	.41***	.51***	.56***	.52***
Girls	.38***	.21	.34*	.54***	.40**
Boys	.60***	.63***	.64***	.59***	.69***
Prealgebra	.46***	.47**	.41*	.54**	.56***
Girls	.23	.13	–.32	.36	.41
Boys	.60***	.76***	.66***	.70***	.70***
Algebra	.55***	.57***	.52***	.62***	.57***
Girls	.50***	.52**	.51**	.59***	.50**
Boys	.62***	.61**	.55**	.66***	.73***

Note. OE = open-ended; MC = multiple-choice; Eff = self-efficacy; Perf = performance.
$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

choice vs. open-ended), performance test format (multiple-choice vs. open-ended), and the interaction between the self-efficacy and performance test formats. In addition, we examined differences across subgroups in the predictive value of self-efficacy by including the interaction between self-efficacy and (a) gender, (b) self-efficacy test format, (c) performance test format, and (d) self-efficacy test format by performance test format (a three-way interaction). We examined the residuals of the multiple regression analyses for heteroscedasticity; plots of the residuals showed the data to be homoscedastic. Tests for nonlinearity were nonsignificant. The model for this analysis was significant ($R^2 = .57$), but the last three interactions described proved to be nonsignificant. Consequently, we tested a reduced model with the main effects, the interaction of the self-efficacy and performance tests formats, and the interaction between self-efficacy score and gender (see Table 3). The reduced model suffered no significant loss in predictive value, $F(6, 320) = 72.39$, $p < .0001$, $R^2 = .57$.

Results of the reduced model showed that the form of self-efficacy assessment was not predictive of performance scores. This finding was consistent with the MANOVA finding of no significant differences between self-efficacy assessments. As expected, the main effect for performance format was predictive; students who took the multiple-choice test scored an average of 3.67 problems higher than did students who took the open-ended performance measure. Both math self-efficacy, $t = 6.36$, $p < .0001$, and algebra level, $t = 13.04$, $p < .0001$, proved predictive of performance. Students enrolled in algebra scored an average of 6.3 problems higher than did students enrolled in prealgebra.

One interesting effect was the significant interaction between gender and self-

TABLE 3
Multiple Regression Model Predicting Mathematics Problem-Solving Performance

Independent variable	Parameter estimate	*SE*	*t*	β	Prob > |*t*|
Intercept	−2.63	2.21	−1.19	.000	.2348
Self-efficacy format	0.56	0.43	1.30	.048	.1933
Performance format	3.63	0.43	8.46*	.309	.0001
Math self-efficacy	0.10	0.02	6.36*	.331	.0001
Level	5.78	0.44	13.04*	.481	.0001
Gender	−7.73	3.14	−2.46*	−.655	.0145
Gender × Self-Efficacy	0.05	0.02	2.55*	.693	.0113

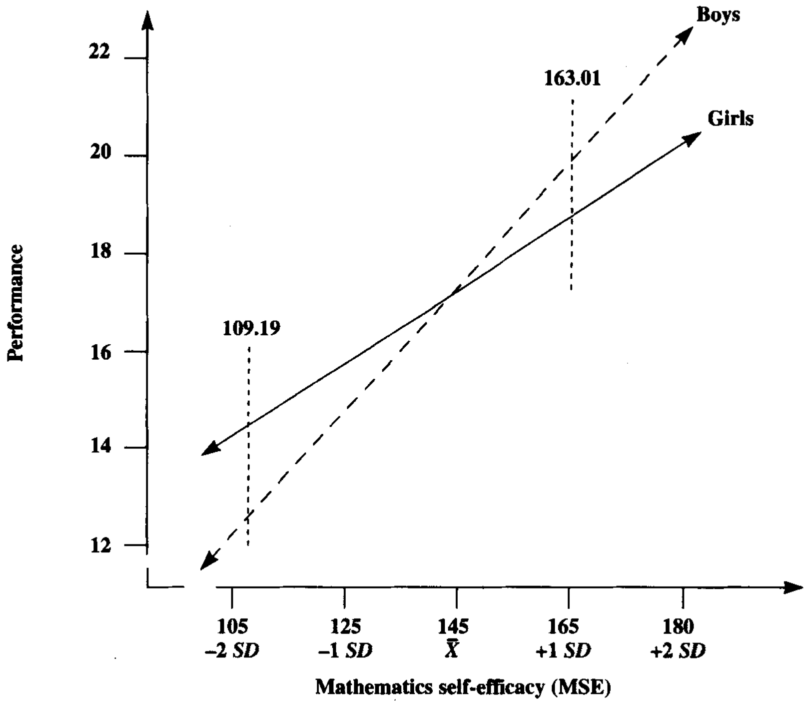
Note. Format denotes multiple-choice versus open-ended response assessment. Level denotes prealgebra (coded 0) versus algebra (coded 1) enrollment. For gender, girls were coded 0 and boys coded 1. For self-efficacy format, the open-ended instrument was coded 0 and the multiple-choice instrument was coded 1. For performance format, the open-ended test was coded 0 and the multiple-choice test was coded 1.

efficacy. That effect was foreshadowed by observed differences in the zero-order correlations (see Table 2). Both the correlations and interaction effect indicate that boys were better predictors of their performance than were girls, but the nature of the interaction itself proved interesting. The nature of the gender by self-efficacy interaction prevented adjusted mean differences in self-efficacy from surfacing (see Figure 2). The disordinal interaction needs to be interpreted in light of the Johnson–Neyman regions of significance, which are bounded at 109.19 and 163.01. This is at z scores of −1.77 and .92. Thus, girls with low self-efficacy performed better than boys with low self-efficacy (below −1.77, representing only 3.84% of the population), whereas boys with high self-efficacy performed better than did girls with high self-efficacy (above .92, representing 17.88% of the population). The latter number seems especially important because close to one fifth of the distribution (17.88%) showed boys outperforming girls at the same level of high confidence. The significant difference in the slopes indicates that the relationship between efficacy and performance was stronger for boys than for girls.

Calibration

The students who were administered the open-ended performance measure had greater bias and lower mean and item accuracy (see Table 1). These differences resulted from the lower performance scores obtained by the OE Eff/OE Perf and MC Eff/OE Perf groups on the open-ended performance measure but without the corresponding lower self-efficacy scores. The MANOVA results revealed significant differences between the students enrolled in algebra and prealgebra for the dependent variables of performance, mean bias, mean accuracy,

FIGURE 2. Interaction of gender and self-efficacy on problem-solving performance.



Note. To plot the regression lines, we used the following formulas: For girls, PERFORMANCE = 5.69 + .08MSE; for boys, PERFORMANCE = −1.20 + .13MSE.

and item accuracy, Wilks's lambda = .63, $F(5, 317) = 37.88$, $p < .0001$ (see Table 4). Consistent with the multiple regression results of the present study, the students enrolled in algebra and prealgebra did not differ in self-efficacy. As expected, algebra students outperformed prealgebra students on both performance measures, and these differences were responsible both for the generally stronger correlations in three of the four groups (see Table 2) and for the higher calibration scores of the algebra students. Both the prealgebra and algebra students expressed confidence that they could solve 80% of the problems, and these judgments did not differ significantly in terms of the strength of their confidence. The analyses revealed no significant gender effects and no interaction effects among group, level, and gender.

In summary, regardless of which self-efficacy instrument was administered, calibration scores were lower for the students subsequently tested with the open-

TABLE 4
Mean Scores for Dependent Variables, by Algebra Level and Gender

Variable	Prealgebra		Algebra		Girls		Boys	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Self-efficacy	142.01	20.60	146.35	19.54	142.96	19.1	146.67	20.98
Multiple choice	140.82	22.52	148.16	16.69	144.22	17.5	146.53	21.40
Open-ended	143.24	18.52	144.42	22.12	141.73	20.6	146.84	20.67
Performance	13.36*	5.11	19.69*	4.95	17.12	5.7	17.32	6.09
Multiple choice	15.33*	4.61	21.52*	3.90	19.00	4.9	19.19	5.50
Open-ended	11.27*	4.82	17.76*	5.21	15.20	5.9	15.27	6.08
Mean bias	1.51*	.81	0.60*	.72	0.91	.92	1.00	.81
Mean accuracy	2.71*	.55	3.38*	.50	3.08	.58	3.17	.65
Item accuracy	15.98*	4.15	21.12*	3.60	18.99	4.51	19.24	4.66

Note. Mean differences reported are the result of MANOVA and follow-ups, with group, gender, algebra level, and all possible interactions as independent variables. No interactive effects of level or gender were found for any variable.

ended performance measure than for those tested with the multiple-choice measure. The algebra students had higher performance scores and better calibration than did the prealgebra students. There were no gender differences for any of the dependent measures.

Discussion

The first objective of this study was to determine whether a traditional, multiple-choice presentation of mathematics problems versus an alternative, open-ended presentation influences students' mathematics self-efficacy judgments. The data analysis demonstrated that the students' reported self-efficacy was not affected by these variations in presentation format. One possible explanation for this result is that the students did not look at the multiple-choice answers when they made their confidence judgments. As we discuss below, it is also possible that, regardless of how their confidence is assessed, students' familiarity with traditional forms of assessment creates a mind set that causes them to base their confidence judgments on the expectation that the performance task will be presented in a traditional, and therefore familiar, format. Supporting this mind set is the fact that students in the two self-efficacy groups were not informed that the format of the performance task might differ from the format of the efficacy task. In other words, when students in the MC Eff/OE Perf group were asked to provide confidence judgments on multiple-choice problems, they were not told that they would be asked to take the subsequent test in an open-ended format. Students in the OE Eff/MC Perf group were also not told of the change in testing format. Had this information been provided, it might have altered confidence judgments. We recommend that replications of the present study include this information as part of the experimental conditions. Nonetheless, it is interesting that confidence judgments predicted task performance as accurately when the formats did not correspond as when they did correspond.

The second objective of the study was to explore the relationship between students' self-efficacy judgments and subsequent performance on traditional versus alternative assessments of mathematical problem solving. The students who took the multiple-choice test received higher scores, likely the consequence of guessing several correct answers. However, the variations in assessment of self-efficacy did not affect the prediction of mathematics performance. Because self-efficacy judgments did not vary across formats, the lower scores obtained on the open-ended performance test created a situation in which discrepancies between judgments and performance were more pronounced for the MC Eff/OE Perf and OE Eff/OE Perf groups, resulting in greater overconfidence and poorer calibration for these groups. In previous research, mathematical problem solving has been assessed by multiple-choice performance measures and researchers have found students to be overconfident about their capability to solve mathematics problems. When performance is assessed with a test that will yield lower scores as a consequence only of the format used, overconfidence and poor calibration will obviously be magnified: Students who express the same confidence level obtain lower performance scores and are thus deemed even more overconfident and more poorly calibrated.

One implication of this finding is that, if open-ended performance tests can be said to provide a more accurate measure of students' actual capability (scores are not inflated by guessing), students may be even less well calibrated and more overconfident about their mathematics capability than researchers have previously reported. Thus, students' self-perceptions of mathematics capability may be even less accurate than has previously been reported.

The poorer calibration and greater overconfidence may also be related to students' familiarity with multiple-choice assessment formats. Students may be so familiar with traditional multiple-choice examinations that their expectation of a future performance task is that it will be multiple choice. This expectation may guide their self-perceptions of capability to the extent that, regardless of how their confidence is assessed, they assume that the examination to follow will be in the format to which they are accustomed and will yield results with which they are also familiar. If this is so, would this expectation vary with changes in assessment that alter the familiar expectations? In other words, if students were made familiar with alternative, open-ended mathematics assessments, would they lower their level of self-efficacy and improve their calibration by virtue of altered perceptions of how they will be tested and how well they will perform? Additional research in this area is required.

Our results confirm that understanding the measurement of self-efficacy and performance is crucial to understanding the relationship between these two variables. Although we found no difference between the two methods of assessing self-efficacy, relations between self-efficacy and performance differed depending on the method of assessing performance. These differences altered the predictive utility of self-efficacy and influenced calibration results. As we have demonstrated, these differences have measurement implications for researchers attempting to assess the relationship between self-efficacy beliefs and related academic outcomes. These measurement concerns must be analyzed further before researchers can make sound generalizations about the strength of the self-efficacy/performance relationship or the accuracy of students' self-perceptions.

Whereas some researchers (Fennema & Hart, 1994; Pajares, 1996a; Wigfield, Eccles, MacIver, Reuman, & Midgley, 1991) have reported that gender differences in mathematics confidence surface in middle school, we did not find them in the present sample. We found, however, that at higher self-efficacy levels, boys were slightly better predictors of their performance than were girls. At lower self-efficacy levels, boys were poorer predictors. Research is needed to better determine the nature of these effects. Pajares (1996a) found no gender differences in the calibration of regular-education middle-school students but discovered that gifted girls were biased toward underconfidence.

Students enrolled in algebra were better calibrated across all measures. This finding is consistent with those of other studies (Pajares, 1996a; Pajares & Kranzler, 1995) of the relationship between calibration and mathematics capability. Pajares (1996a) reported that middle-school gifted students are better calibrated than are regular-education students. Similarly, we found that the more capable students in the present sample were better judges of their capability. In general, we concur with the recommendation of Pajares and Kranzler (1995) that instructional intervention is needed to help students better understand what they know and do not know so that they can more effectively deploy appropriate cognitive strategies during mathematical problem solving. Interventions should be particularly appropriate for students at lower levels of academic achievement. It seems likely that a productive intervention would be to vary the form of assessment and to familiarize students with each form. These recommendations are particularly pertinent to states and school districts that are moving toward new forms of assessment.

The present findings are based on a sample of regular-education eighth-grade students enrolled in higher level mathematics courses. We recommend that the findings be tested using students enrolled in other levels of mathematics and in other grades. In addition, we acknowledge that the strength of the relationship between self-efficacy and performance may be influenced by the correlated specifics that can result from the use of the same items to measure both constructs. Marsh, Roche, Pajares, and Miller (in press) have cautioned that using identical self-efficacy and performance indexes in an effort to closely match belief and criterion may lead to positively biased estimates of effects from self-efficacy to performance outcomes. Thus, researchers are encouraged to use similar rather than identical items or tasks to assess self-efficacy beliefs and performance criteria or to use structural equation modeling analyses to sift out the bias that might result from correlated specifics.

NOTE

We wish to express our gratitude to Mel Lucas, Professors Tim Urdan and Margaret Johnson, and especially to Gio Valiante for their valuable assistance on this project. In addition, we thank the consulting editors and the executive editor of *The Journal of Experimental Education* for the many valuable comments and suggestions that served to strengthen the final manuscript. This study was funded by a grant from the Florida Educational Research Council, Inc.

Address correspondence to Frank Pajares, Department of Educational Studies, Emory University, Atlanta, GA 30322.Tel: (404) 727-1775. Fax: (404) 727-2799. E-mail: mpajare@emory.edu.

REFERENCES

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

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Betz, N. E., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior, 23,* 329–345.

Collins, J. L. (1982, March). *Self-efficacy and ability in achievement behavior.* Paper presented at the meeting of the American Educational Research Association, New York.

Fennema, E., & Hart, L. E. (1994). Gender and the JRME. *Journal for Research in Mathematics Education, 25,* 648–659.

Fennema, E., & Sherman, J. A. (1978). Sex-related differences in mathematics achievement and related factors: A further study. *Journal for Research in Mathematics Education, 9,* 189–203.

Hackett, G. (1985). The role of mathematics self-efficacy in the choice of math-related majors of college women and men: A path analysis. *Journal of Counseling Psychology, 32,* 47–56.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education, 20,* 261–273.

International Association for the Evaluation of Educational Achievement (1985). *Second Study of Mathematics: Technical Report IV.* Washington, DC: Author.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Lapan, R. T., Boggs, K. R., & Morrill, W. H. (1989). Self-efficacy as a mediator of investigative and realistic general occupational themes on the Strong–Campbell Interest Inventory. *Journal of Counseling Psychology, 36,* 176–182.

Marsh, H. W., Roche, L. A., Pajares, F., and Miller, D. (in press). Item-specific efficacy judgments in mathematical problem-solving. *Contemporary Educational Psychology.*

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high stakes environment. *Educational Measurement: Issues and Practice, 12,* 9–15.

Miller, M. D., & Seraphine, A. E. (1993). Can test scores remain authentic when teaching to the test? *Educational Assessment, 1,* 119–129.

Pajares, F. (1996a). Self-efficacy beliefs and mathematical problem solving of gifted students. *Contemporary Educational Psychology, 21,* 325–344.

Pajares, F. (1996b). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66,* 543–578.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 26,* 426–443.

Pajares, F., & Miller, M. D. (1994). The role of self-efficacy and self-concept beliefs in mathematical problem-solving: A path analysis. *Journal of Educational Psychology, 86,* 193–203.
Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics outcomes: The need for specificity of assessment. *Journal of Counseling Psychology, 42,* 190–198.
Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33–40.
Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9,* 321–332.
Schraw, G., Dunkle, M. E., Bendixen, L. D., & DeBacker Roedel, T. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87,* 435–444.
Schraw, G., & DeBacker Roedel, T. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22,* 63–69.
Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18,* 445–463.
Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review, 1,* 173–208.
Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26,* 207–231.
Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29,* 332–329.
Wigfield, A., Eccles, J., MacIver, D., Reuman, D., & Midgley, C.(1991). Transitions at early adolescence: Changes in children's domain specific self-perceptions and general self-esteem across the transitions to junior high school. *Developmental Psychology, 27,* 552–565.
Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46,* 41–47.
Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B, 21,* 396–399.
Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan, 74,* 444–456.
Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.