Journal of Adult Education
Volume 38, Number 1, 2009

Gender Differences in Mathematics Self-Efficacy and Back Substitution in Multiple-Choice Assessment

K. Shane Goodwin
Lee Ostrom
Karen Wilson Scott

Abstract

A quantitative observational study exploring the relationship of gender to mathematics self-efficacy and the frequency of back substitution in multiple-choice assessment sampled undergraduates at a western United States parochial university. Research questions addressed: to what extent are there gender differences in mathematics self-efficacy, as demonstrated on multiple-choice test items; and to what extent are there gender differences in the frequency of employing back substitution as an informed guessing strategy on multiple-choice test items? Instruments were (a) a representative multiple-choice test algebra equation, and (b) a mathematics self-efficacy survey accompanying a standardized multiple-choice algebra examination. While results revealed no significant gender differences in mathematics self-efficacy or back-substitution strategy, findings concerning testwise strategies application and learner performance accuracy can benefit educators.

Introduction

Although multiple-choice examinations are carefully and objectively scored, they can introduce significant variability and measurement error

K. Shane Goodwin is Professor, Mathematics, Brigham Young University-Idaho, Rexburg, ID. Lee T. Ostrom is Associate Professor, Adult, Career, and Technology Education, University of Idaho, Idaho Falls, ID. Karen Wilson Scott is Associate Professor, Human Resource Training and Development, Idaho State University, Pocatello, ID.

due to both random and informed guessing. Informed guessing is a part of a broader group of testwiseness skills – the skills that aid in selecting the correct answer without actually knowing the material being tested (Cronbach, 1984; Hoffman, 1962; Owen, 1985; Payne, 2003). Sarnacki (1979) reviewed many studies that found how the application of testwiseness skills on multiple-choice tests emerges throughout the disciplines and spans all ages from preschool children to adults.

In undergraduate mathematics, multiple-choice examinations are often utilized in lower division courses as well as for entry-level placement examinations required for program admittance. Zachai (1995) details some of the struggles adult learners have with these competency tests such as test anxiety, insufficient test-taking skills, and lack of practice with speeded tests. Thus, the approaches of improving test-taking skills combined with refreshing mathematical skills such as algebra are critical for success. The psychological and cognitive adhesive that bonds these two approaches together is mathematics self-efficacy – the specific self-assessed belief in one's own capability of solving mathematical problems and tasks successfully (Hackett & Betz, 1989).

Along with testwiseness principles that introduce subjectivity into the objective world of multiple-choice testing, researchers, over recent decades, have vigorously debated gender bias (e.g., Beller & Gafni, 2000; Haladyna, 1994; Hassmén & Hunt, 1994; Quereshi, 1974; Skinner, 1983). Few academic areas provide as rich a forum in the gender difference debate as mathematics (e.g., Fennema & Leder, 1990; Spelke, 2005).

The problem addressed by this study was twofold: First, while the literature is robust in terms of gender bias studies of the multiple-choice format for examinations, the potential gender differences in self-efficacy and testwiseness for adult learners engaged in undergraduate mathematics courses are not as well understood. Second, gender studies often fail to statistically control for variables such as mathematical background (Eagly, 1995; Haladyna, 1994; Parlee, 1981) or place adequate emphasis on effect sizes and statistical power (Cohen, 1992, McLean & Ernest, 1998, Vacha-Haase, 2001). Undergraduate students at a private four-year university in southeastern Idaho were sampled as part of a quantitative observational study exploring the relationship of gender to both mathematics self-efficacy and the frequency of back substitution in multiple-choice assessment. Two research questions were addressed:

(a) To what extent are there gender differences in mathematics self-efficacy, as demonstrated on multiple-choice test items? (b) To what extent are there gender differences in the frequency of employing back substitution as an informed guessing strategy on multiple-choice test items? As the observational study was conducted in two parts, this paper will discuss the two questions separately.

Mathematics Self-Efficacy and Gender Differences Background

The higher education backdrop for this research can be noted by the dramatic changes observed in gender balance in recent years. Kindlon (2006), synthesizing the National Center on Educational Statistics for the 2004-2005 academic year, states, "Women earned 62 percent of all associate's degrees, 59 percent of all bachelor's degrees, 60 percent of all master's degrees, 48 percent of doctorates, and 51 percent of professional degrees" (p. 8). Using data from the U.S. Department of Labor Statistics in 2003, Meece, Glienke, and Burg (2006) state that:

> Over the last three decades, unprecedented changes in women's level of educational participation and occupational status have been observed. For the first time in U.S. history, women are earning more college degrees than men, and they exceed men in many fields of study including psychology, accounting, and health-related professions. (p. 352)

These shifts in the gender balance of degrees motivate further exploration of the role that gender plays in higher education assessment.

Summarizing historically, Zachai points out that from the 1950s through the 1970s, women were avoiding higher mathematics and entered college unprepared for the challenges and rigor of college courses. When faced with gatekeeping examinations for admission or placement into programs, female adult learners "were at a disadvantage" (p. 8). Kindlon (2006) proposes that the prevailing wisdom of that era was that as girls move into adolescence, they go through a loss of confidence in themselves and especially in their ability to assimilate math and science. Also, more women reported math anxiety and mathematics courses avoidance than men (Hembree, 1990; Tobias 1978).

In reviewing the literature base on gender and developmental college-level mathematics, Stage and Kloosterman (1995, p. 294) report

that fewer than half of students taking remedial courses pass on the first try and a disproportionate number of those who fail are women and minorities. Their study addressed the question of how self-efficacy beliefs relate to success in mathematics and were able to show that this correlation was statistically higher for female students than for male students.

As non-traditional students progress through the rigors of a post-secondary degree, they must grapple with the challenges of an extensive review of algebraic skills from their past schooling and of their individual mathematics self-efficacy—the specific self-assessed belief in one's own capability of solving mathematical problems and tasks successfully (Hackett & Betz, 1989). A personal confidence in acquiring mathematical skills may be considered a more generalized human trait when compared with the specificity of mathematics self-efficacy. Bandura (1997) comments that:

> Confidence is a nondescript term that refers to strength of belief but does not necessarily specify what the certainty is about. I can be supremely confident that I will fail at an endeavor. Perceived self-efficacy refers to belief in one's power to produce given levels of attainment. A self-efficacy assessment, therefore, includes both the affirmation of a capability and the strength of that belief. (p. 382)

The role of gender and its relationship to self-concept, self-efficacy, and tenacity in solving a challenging multiple-choice problem is of interest to a wide variety of researchers (e.g., Bandura, 1997; Eccles, 1983; Gwilliam & Betz, 2001; Hackett & Betz, 1989; Pajares, 2005; Pajares & Miller, 1994; Randhawa, 1994; Sadker & Sadker, 1994; Seegers & Boekaerts, 1996). It is in the area of mathematics where we see even more emphasis placed in self-efficacy studies, perhaps because of the valued role that mathematics plays in academia, high-stakes assessments for admission and scholarships, and the filtering of students in highly technical and specialized careers (Pajares, 2005).

Method

In the first part of this observational inquiry we addressed the research question: To what extent are there gender differences in mathematics self-efficacy, as demonstrated on multiple-choice test

items? The sampled participants of our study were male and female undergraduate students taking a lower division mathematics course in intermediate algebra, quantitative reasoning or college algebra at Brigham Young University-Idaho. BYU-Idaho is a private, religious, four-year university in southeastern Idaho with over 20,000 students attending throughout the calendar year (unduplicated head count) and with a maximum of approximately 13,000 students attending at one time during any one of the three regular semesters. In terms of student body composition, students come from all 50 states and over 55 different nations with married students comprising approximately 26% of the student body. Approximately 40% of the undergraduates have served church missions, thus taking them out of higher education usually for two years. They return with a higher level of maturity and characteristics that are more closely associated with those of adult learners, or non-traditional students.

In order to avoid the confounding variable of self-reported data, human assurance approval allowed, for the purposes of the study, linkage of the demographic information from the university student records system to the survey results of each participant. To statistically control for background variables, which gender studies sometimes fail to do, (see Eagly, 1995; Haladyna, 1994; Parlee, 1981), we gathered data reflecting the independent variables of gender, marital status, missionary status, current mathematics course, GPA, ACT scores (both composite and sub-scores), cumulative college credits, and age.

The sampling design entailed two phases of data collection: (a) a brief in-class questionnaire requiring students (N = 1028) to solve one challenging algebraic equation and then answer two follow-up questions regarding strategy used and confidence level, and (b) a thorough mathematics self-efficacy survey that paralleled 30 problems from a standardized unit examination the intermediate-algebra students (N = 139) were required to take as part of their semester course. Of the undergraduates who completed the in-class questionnaire, 55% were enrolled in a quantitative reasoning course, 25% in college algebra, and 20% in an intermediate algebra. Only the intermediate algebra students were invited to participate later in the semester in the more in-depth mathematics self-efficacy survey. In both phases of data collection, participants reported their mathematics self-efficacy correlating with each test answer via a six-point Likert scale (1 = Random Guess, 2 =

Very Unsure, 3 = Somewhat Unsure, 4 = Somewhat Sure, 5 = Very Sure, And 6 = Certain).

Results

After gathering demographic data for both the undergraduates completing the in-class survey as well as those taking the standardized intermediate algebra test, we tested for gender differences using a non-directional t-test for independent sample means. These *t*-tests revealed no statistically significant differences in cumulative credits, GPA, or ACT composite scores. Men, however, did have statistically significant higher ACT Mathematics sub-score means (M_{male} = 20.4 to M_{female} = 19.7, with a *P*-value = .005 for the students completing the in-class questionnaire and M_{male} = 19.4 to M_{female} = 17.9, with a *P*-value = .007 for the students taking the intermediate algebra standardized test).

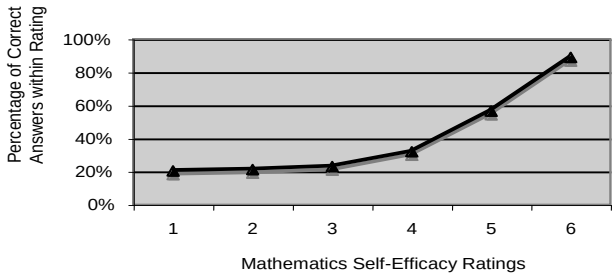
The sampled N = 1,028 participants, who completed the in-class questionnaire, represented a 94% response rate in the first phase. Female students made up n = 539 or 52.4% of the sample and male students made up n = 489 or 47.6%. In the second phase of data collection involving only the intermediate algebra course, students making self-efficacy ratings on their actual intermediate algebra examination constituted the sample of N = 139 represented a response rate of 67%. The female sub-sample size was n = 64 or 46.0%, and the male sub-sample size was n = 75 or 54.0%. This lower response rate was anticipated due to the greater commitment required by completing the self-efficacy survey for all 30 questions of a live algebra test administered at the university's testing center.

Ben-Shakhar and Sinai (1991) point out that in many studies, participants either volunteered to take tests or were paid. They suggest that realistic testing situations might provide more honest motivation in answering test items. Therefore, the present study assessed an actual testing situation. Participants (N = 139) in the university testing center were encouraged to assess their level of confidence in each answer upon completion of each question rather than waiting until the end of the test and relying on memory. They were also reminded that if the survey became too distracting during their testing experience that they could simply refrain from completing the survey question. Although the sample size was small in comparison to that of the in-class questionnaire, we

concluded that it added to the validity of the data and the inferences made therein.

Reliability of the data from both phases of collection was carefully analyzed through several descriptive and inferential methods. Although one cannot know the seriousness that any one particular student gave to the survey, in the aggregate it is possible to assess the consistency and seriousness of the participants. An example of consistency in the in-class questionnaire data, Figure 1 depicts a trend between how well the students performed on the algebraic problem and what mathematics self-efficacy rating they selected on the survey. If the data were unreliable, we would expect somewhat of a random pattern to occur between correctness and ranking. On the other hand, if the data were reliable, we would expect to see an upward trend in percentage of correct answers the higher they ranked their mathematics self-efficacy. Approximately 20% of those students who selected the first ranking (random guess) did get the question correct. This was not surprising, considering the multiple-choice question contained five options from which to choose (a, b, c, d, or e) and therefore, if they were truly guessing randomly, we would expect a performance close to 20%.

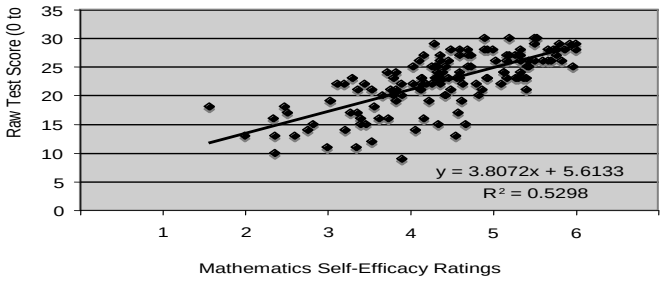
Figure 1: Trend of Correct Answer Percentage Over the Self-Efficacy Ratings



An example of reliability checking for the second phase of data collection was a comparison of the student's individual score with his or her mathematics self-efficacy mean of the 30 problems. As Figure 2

illustrates, there is a relatively strong correlation (r = 0.73) between the student's raw test scores in intermediate algebra and his or her corresponding mathematics self-efficacy mean on the test. A linear regression significance test produced $t(138) = 4.06$, a coefficient of determination of $R^2 = 0.5298$ and $p < .0001$. Thus, about 53% of the variation in the student's raw test score can be explained by the least-squares model.

Figure 2: Scatter Plot of Test Scores Versus Self-Efficacy Ratings



Regarding the first research question of this study, "To what extent are there gender differences in mathematics self-efficacy, as demonstrated on multiple-choice test items?" an independent samples t-test for population means was conducted on both the in-class questionnaire and the testing center data. From the in-class questionnaire data, we observed that male undergraduates had a slightly higher mean (M = 3.02) in mathematics self-efficacy, but it was not statistically significant with $t(1026) = .927$, p-value = .354 (assuming equal variances). Cohen's measure of effect size obtained was $d = .057$, therefore not providing sufficient sample evidence to support the claim of a statistically significant gender difference in mathematics self-efficacy. Using multiple regression with SPSS 14.0, we wanted to address the issue of whether men and women have different mathematics self-efficacy means after controlling for various academic and personal demographics. After checking for multicollinearity in the model and using the backward elimination process (eliminating the least statistically

significant to the most—one by one) to remove non-significant variables from the regression model, we obtained results as indicated in Table 1. The only significant variables to be retained in the model were cumulative credits and the ACT Mathematics sub-score. We theorized that the negative coefficient in the model was attributable to the notion that, the more cumulative credits students had earned, the longer they had postponed their graduation requirement for mathematics, therefore the lower their mathematics self-efficacy ranking due to a greater likelihood of math phobia. In summary, even when statistically controlling for demographic background, gender did not play a statistically significant role in the regression model of mathematics self-efficacy.

Table 1: Multiple Regression of In-class Questionnaire After Backward Elimination

	Unstandardized Coefficients				
Model Coefficients	*B*	*SE*	*t*	*df*	*P-value*
(Constant)	1.991	.253	7.854	1	.000
Gender	.004	.087	.046	1	.963
Cumulative Credits	-.005	.002	-2.687	1	.007
ACT Mathematics	.058	.012	4.998	1	.000

In the second phase of data collection, using the mathematics self-efficacy survey administered alongside an actual 30-question intermediate algebra examination in the testing center, similar results yielded no statistically significant gender difference with $t(138) = -.544$, $p = .587$ (with assumed equal variances based on Levene's Test of $p = .173$). Cohen's effect size was $d = .092$.

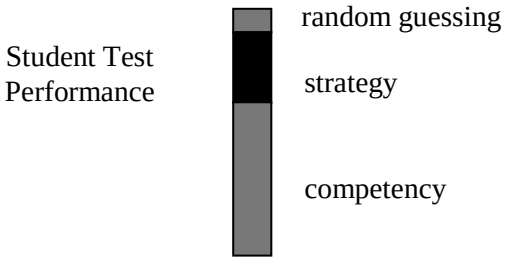
Back Substitution and Gender Differences Background

Despite careful and objective scoring, multiple-choice examinations can introduce significant variability and measurement error due to both random and informed guessing. Fagley (1987) indicates in a study involving 62 university students taking a difficult 70-question history

examination, that about 16% of the variance in tests scores of the participants was attributable to testwiseness. It is this variability in test performance that serves as an important backdrop to the second research question of our study regarding a testwise strategy known as back substitution—a specific strategy in which students substitute the test item options into an equation until the correct answer is discovered amid the distractors (see Bridgeman, 1992; Katz, Bennett, & Berger, 2000). Back substitution is also referred to as *working backwards*, *reverse engineering*, or *plugging it in*. Bridgeman (1992) calls attention to it in his study of contrasting multiple-choice with open-ended item formats. He states "an algebraic problem such as $2(x + 4) = 38 - x$ becomes a much simpler arithmetic problem if the examinee can just substitute the values of x given in the answer choices until the correct value is found" (p. 253).

Figure 3 displays a simple model partitioning a student's test score into three segments. The percentage of the score that is honestly earned through applying correct procedures and recall was operationally defined as *competency*, the percentage dedicated to testwiseness or informed guessing as *strategy*, and the remaining portion as *random guessing*. For students relying on testwise strategies rather than competency, one of the possible strategies applied is back substitution.

Figure 3: Partitioning of Student Test Performance



Over the decades, researchers have argued that there is evidence of potential gender biases in the multiple-choice format. For example, Ben-Shakhar and Sinai (1991) determined that on average, females tend to omit more problems than males. This occurred even when guessing penalties were not applied and permissive instructions were given to encourage guessing whenever examinees were unsure about their answers. In their survey of gender bias research, Hassmén and Hunt (1994) found in their literature review that female examinees tend to rely more on recall, while male examinees rely more on problem solving strategies when taking mathematical tests. Hassmén and Hunt also summarized that female students on average:

> Are less able than males to take advantage of extraneous clues, are more inclined to rely on recall instead of problem solving, leave more questions unanswered than males, are more inclined to change the answers they do give, and consider multiple-choice tests as more problematic/ challenging than males do. (p. 152)

Pascale (1974) was able to show that women were twice as likely to change their original answers than men and yet men did better than women on the questions that they did change. In a later study with psychology undergraduates, Skinner (1983) found similar results. Bridgeman and Lewis (1994) indicate that studies in Ireland, Australia, and Great Britain have consistently shown that men may have an advantage over women when it comes to the multiple-choice style of testing. Beller and Gafni (2000) theorize that a potential reason for this statistical interaction of gender and format might be due to the different risk-taking tendencies of girls and boys and their different strategies in response to a multiple-choice test item.

Literature in recent years also suggests that there are many other perspectives to consider in the spirited debate among academic researchers. Defending the multiple-choice format, Haladyna (1994) passionately calls for more gender research to be conducted using control variables such as prior mathematical experience, writing skills, interest, motivation, test-taking skills, and so forth. Haladyna, Downey, and Rodriguez (2002) comment that “critics have often noted that item writing is an immature science” (p. 309), and studies show that students score higher after having completed a test-taking course or workshop (Sarnacki, 1979; Scruggs & Mastropieri, 1992), which also impacts the variability of the test scores due to testwiseness.

Method

In the second part of this observational inquiry, we engaged the research question: To what extent are there gender differences in the frequency of employing back substitution as an informed guessing strategy on multiple-choice test items? The sampling design addressing research question 2, also took advantage of the first phase of data collection: a brief in-class questionnaire requiring students (N = 1028) to solve one challenging algebraic equation and then answer two follow-up questions regarding strategy used and confidence level. Of the 1028 participants responding to the survey, N = 1005 completed the back substitution portion of the in-class questionnaire.

Results

In order to quantify any potential gender differences in the application of the back-substitution strategy, two separate chi-square tests of independence were calculated. The first was a check for differences in the male-female distribution of equation-solving strategies. Figure 4 illustrates just how similar the sampled men and women were in their selection of strategy for solving the equation from the questionnaire. Students were asked upon completing the equation-solving problem, which of the strategies (random guessing, back-substituting, doing algebra without checking, or doing algebra with checking) had they employed in coming to their final answer.

Figure 4*:* Gender Comparison Between the Four Equation-Solving Strategies

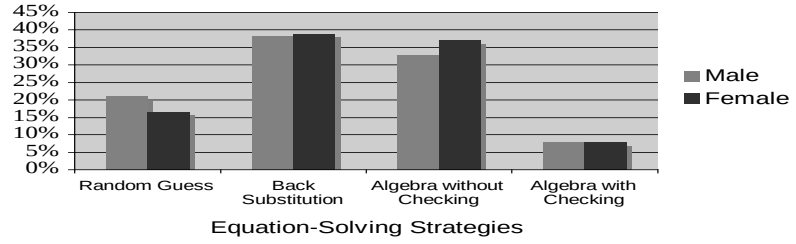


Table 2 contains the actual survey count data upon which the relative percentages were calculated among the four strategies. Results obtained after a chi-square test for independence were $\chi^2(3, N = 1{,}005) = 4.021$, $p = .259$, suggesting that the choice of strategy was independent of gender. That is, there was not a statistically significant difference in the frequency of application of these four equation-solving strategies between men and women.

Table 2: In-Class Questionnaire Results of Equation-Solving Strategy and Gender

Strategy	Male Participants (n = 477)	Female Participants (n = 528)
Random Guess	100	87
Back Substitution	183	205
Algebra without Checking	156	195
Algebra with Checking	38	41

The second check for gender difference involved directly analyzing the back-substitution proportions in an isolated fashion. From the in-class questionnaire, 38.4% of men back substituted the equation while 38.8% of the women did so. To test the proportions, a chi-square test of independence addressed the gender question with results of $\chi^2(1, N = 1{,}005) = .022$, $p = .881$. Cramer's Phi for effect size was $\phi_C = 0.005$ suggesting there was no statistical evidence supporting the claim of gender difference in the employment of the back-substitution strategy.

To round out the analysis on the second research question, it was necessary to statistically control for the demographic and academic variables of the study by using multiple logistic regression which takes into account the dichotomous nature of the response variable of back substitution. Table 3 shows the results after a process of backwards elimination of variables as was done in the multiple regression for the first research question. It was observed that ACT Mathematics and Class (intermediate algebra, quantitative reasoning, or college algebra) remained as statistically significant coefficients. Gender, however, was

not statistically significant in the final logistic regression model of back substitution even after controlling for demographic variables delineated above.

Table 3: Multiple Logistic Regression After Backward Elimination

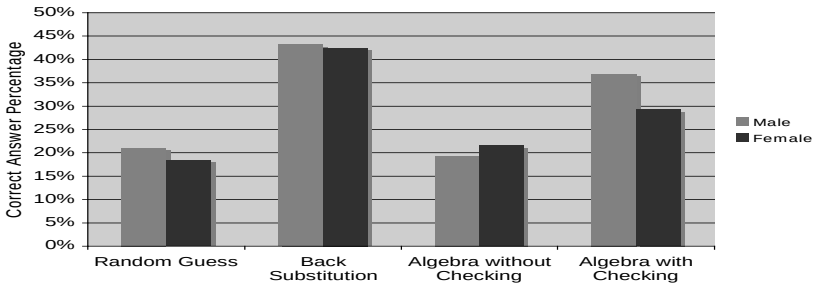
Model Coefficients	*B*	*SE*	Wald's Chi-square	*df*	*p*
(Constant)	-2.494	.432	33.291	1	.000
Gender	.145	.133	1.180	1	.277
Class			24.985	2	.000
ACT Mathematics	.080	.018	19.234	1	.000

Additional Analysis

A few more comparisons from the data analysis are noteworthy even though they are tangential to the study's research questions. Referring back to Figure 4 and the notion of guessing on multiple-choice test items, of the 477 men who responded to the strategy question, 21.0% said they had randomly guessed and of the 528 women, 16.5% said the same. A chi-square test on those two proportions obtained the results of $\chi^2(1, N = 1{,}005) = 3.33$, $p < .068$. This could be considered statistically significant at the $\alpha = .10$ level, suggesting perhaps, that men tend to randomly guessing (in this context) at a higher rate than women when given the opportunity.

Overall, the similarities between genders were again strikingly apparent when contrasting performance on the equation-solving test item and the strategy chosen. It is worth pointing out that on the in-class questionnaire, 30.7% of the male participants correctly identified the answer to the equation compared to 30.4% of the females—also not statistically significant with $p = .914$. Figure 5 illustrates the four different strategies and the percentage of correct answers within gender.

Figure 5: Correctness Percentages by Gender Based On Strategy Used



The accuracy on the in-class questionnaire did not dramatically differ between genders on three of the four basic equation-solving strategies as observed in Figure 5. The exception was the method of doing both algebra and checking the work. Of the men choosing to solve with algebra and check the solution, 36.8% were correct. Of the women choosing this same strategy and checking the solution, 29.3% were correct. These two percentages were statistically significant with results of $\chi^2(1, N = 1{,}005) = 7.86$, $P = .005$. In terms of overall accuracy on the equation-solving test item (combining genders), the four strategies ranked from low to high as follows:

- 19.8% for the random guess strategy
- 20.6% for the algebra *without* checking strategy
- 32.9% for the algebra *with* checking strategy
- 42.8% for the back-substitution strategy

Finally, when combining self-efficacy ratings 1 through 3 as Unsure and ratings 4 through 6 as Sure, Table 4 depicts the proportion breakdown according to scenarios and gender. There were no statistically significant gender differences detected in any of the four scenarios but rather more descriptive weight was given to support gender similarities.

Table 4: Comparing Correctness Scenarios for Mathematics Self-Efficacy and Gender

Scenarios	Male	Female
Sure and Correct (Confident)	16.7%	16.0%
Sure and Wrong (Over-Confident)	19.1%	17.1%
Unsure and Correct (Under-Confident)	14.0%	14.3%
Unsure and Wrong (Anxious)	50.2%	52.5%

Conclusions

Having self-confidence in general and high self-efficacy in particular, could make a substantial difference for the adult learner in undergraduate mathematics. Better understanding the relationship of gender with this idea of mathematics self-efficacy would help teachers to be more effective in their classroom management as well as assessment. Zimmerman and Schunk (2003) suggest:

> Teachers who consider their students' self-efficacy beliefs, goal setting, strategy use, and other forms of self-regulation in their instructional plans not only enhance students' academic knowledge, but they also increase their students' capability for self-directed learning throughout their life span. (p. 452)

Knowing the level at which the adult learner performs on a mathematics examination is vital for the instructor, but it is equally important to better understand the student's perceptions of his or her own performances and weaknesses. With multiple-choice assessment, an educator can know much about the performance of a class in the aggregate, but on the level of any given student, the assessment of student mastery of material is much cloudier due to random and informed guessing variability. Although item analysis of multiple-choice tests provides rich, detailed statistics on both student and class performance, the measurement of competency on any specific test item for any specific student is still fraught with variability.

Willingham and Cole (1997) admirably synthesize multiple gender-based studies of fairness in assessment and address directly the multiple-choice format controversy. In their review of 12 well-known studies

dealing with constructed-response versus multiple-choice format issues, none showed a statistically significant format effect (p. 276), thus weakening the argument for gender differences. However, they raise the issue of construct effect being a much more likely candidate for gender differences in performance. This effect has been discussed in the literature (see Frederiksen, 1984). Construct effect essentially refers to the possibility that, since multiple-choice items tend to test an examinee on more discrete and decontextualized knowledge compared to constructed-response items, perhaps there is an inherent gender bias regarding *what it tests* as opposed to *how it tests*.

In terms of testwiseness and gender differences in the context of the adult learner, this study did not reveal any statistically significant difference in the application of back substitution by male and female students. Had it have been otherwise, the notion of gender bias in this context of test strategy would be discomforting for both teachers and learners. If either gender were found to be taking more of an advantage of the multiple-choice format, then the claim of fair assessment could be called into serious question.

The ability to select correct responses to questions without understanding the material can lead to a false sensation of topic mastery. Back substitution is just one example of testwiseness skills that students can develop over their academic careers and yet still not fully understand the mathematics being tested. Multiple-choice test items will always be prone to the variability of measurement error due to random and informed guessing, but this study perhaps indicates that this format vulnerability is most likely equalized between the genders.

Because random or informed guessing potentially lurks behind each correct answer in the multiple-choice format, there is a need for an enhanced understanding of student self-efficacy in test taking. Hassmén and Hunt (1994) argue “it is not enough that a person can recognize correct answers on a test” (p. 158). They submit that teachers or even employers who use competency assessments need to be aware of the examinee’s confidence to help clear up misconceptions on essential topics and pinpoint a lack of confidence that corresponds to correct answers. These potential relationships between confidence, student demographics, and testwiseness add insights to the ongoing debate of assessment formats and to the psychology of the test-taking adult learner of entry-level undergraduate mathematics.

Claims by Kindlon (2006) and Pajares (2005) suggest the gender gap in confidence seems to be narrowing. Eagly (1995) suggests that there is a popular argument that most gender differences tend to be small and that there is a continuum of gender differences as opposed to a the simple yes or no answer. This paradigm seems to ring true considering the findings of this present study. If what Eagly states is true about a continuum of gender difference, rather than a black or white—yes or no dichotomy, then this present study has revealed more similarities than differences and all for the better. Undoubtedly, the debate will continue into future research of gender-format interactions. However, there may be a converging of evidence.

In the spirit of parsimony, it is noteworthy that there was neither a statistically significant gender difference found in mathematics self-efficacy, nor a statistically higher rate for either men or women in employing the back-substitution strategy. Were that the case, we would be even more concerned about the potential for gender disparity and bias in multiple-choice testing for mathematics. We believe this study not only adds to the empirical debate of fairness in this method of assessment that permeates our society, but also contributes to the literature of testwise strategy application and learner performance accuracy. Moreover, we feel gender similarities are just as essential to understand as are gender differences when it comes to the arena of fair assessment in adult education. While more detailed statistical inquiry is needed to explore these gender issues, we believe these findings offer additional support to educators who advocate the use of well-prepared multiple-choice tests while retaining gender equity in their assessment of adult learning in undergraduate mathematics.

References

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

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles, 42*, 1-21.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 23-35.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29,* 253-271.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31,* 37-50.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cronbach, L. J. (1984). *Essentials of psychological testing,* (4th ed.). New York: Harper & Row.

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist, 50*, 145-158.

Eccles, J. (1983). Expectancies, values, and academic behavior. In J. T. Spencer (Ed.), *Achievement and achievement motivation* (pp. 75-146). San Francisco: W. H. Freeman.

Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology, 79,* 95-97.

Fennema, E., & Leder, G. C. (1990). *Mathematics and gender.* New York: Teachers College Press.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Gwilliam, L. R., & Betz, N. E. (2001). Validity of measures of math- and science-related self-efficacy for African Americans and European Americans. *Journal of Career Assessment, 9*, 261-281.

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.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

Hassmén, P., & Hunt, D. P. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurement, 31,* 149-160.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. Journal for *Research in Mathematics Education, 21,* 33-46.

Hoffman, B. (1962). *The tyranny of testing.* New York: Crowell-Collier.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement, 37,* 39-57.

Kindlon, D. (2006). *Alpha girls: Understanding the new American girl and how she is changing the world.* New York: Rodale.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*, 15-22.

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*, 351-373.

Owen, D. (1985). *None of the above: Behind the myth of scholastic aptitude.* Boston: Houghton Mifflin Company.

Pajares, F. (2005). Gender differences in mathematics self-efficacy beliefs. In A. Gallagher & J. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach.* (pp. 294-315). New York: Cambridge University Press.

Pajares, F., & Miller, M. D. (1994). The role of self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Educational Psychology, 86*, 192-203.

Parlee, M. B. (1981). Appropriate control groups in feminist research. *Psychology of Women Quarterly, 5*, 637-644.

Pascale, P. J. (1974). Changing initial answers on multiple-choice achievement tests. *Measurement and Evaluation in Guidance, 6*, 236-238.

Payne, D. A. (2003). *Applied educational assessment* (2nd ed.). Belmont, CA: Wadsworth/Thompson Learning.

Quereshi, M. Y. (1974). Performance on multiple-choice tests and penalty for guessing. *Journal of Experimental Education, 42*, 74-77.

Randhawa, B. S. (1994). Self-efficacy in mathematics, attitudes, and achievement of boys and girls from restricted samples in 2 countries. *Perceptual and Motor Skills, 79*, 1011-1018.

Sadker, M., & Sadker, D. (1994). *Failing at fairness: How America's schools cheat girls.* New York: Charles Scribner's Sons.

Sarnacki, R. E. (1979). An examination of testwiseness in the cognitive test domain. *Review of Educational Research, 21*, 252-279.

Scruggs, T., & Mastropieri, M. (1992). *Teaching test-taking skills: Helping students show what they know.* Purdue University: Brookline Books.

Seegers, G., & Boekaerts, M. (1996). Gender-related differences in self-referenced cognitions in relation to mathematics. *Journal for Research in Mathematics Education, 27*, 215-240.

Skinner, N. (1983). Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology, 10*, 220-222.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? *American Psychologist, 60*, 950-958.

Stage, F. K., & Kloosterman, P. (1995). Gender, beliefs, and achievement in remedial college-level mathematics. *Journal of Higher Education, 66*, 294-311.

Tobias, S. (1978). *Overcoming math anxiety.* New York: Norton.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61,* 219-224.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Zachai, J. (1995). Adult learners' math self-concept as a barrier to passing California State University's entry level mathematics (ELM) Test. *ProQuest Digital Dissertations*, (UMI No. 9928134).

Zimmerman, B. J., & Schunk, D. H. (2003). Albert Bandura: The scholar and his contributions to educational psychology. In B. J. Zimmerman & D. H. Schunk (Eds.), *Educational psychology: A century of contributions.* (pp. 431-457). Mahwah, N.J.: Lawrence Erlbaum Associates.