Developing And Validating The Females In Mathematics Scale, FIMS

Article · September 2010

3 authors:

Katherine Picho

Jerome Katrichis
University of Hartford

D. Betsy Mccoach
University of Connecticut

Developing and Validating the Females In Mathematics Scale (FIMS)

Katherine Picho
University of Connecticut, USA

Jerome M. Katrichis
University of Hartford, USA

D. Betsy McCoach
University of Connecticut, USA

Abstract

Over the years, research on teacher bias towards females in mathematics has yielded different results, which the study attributes to the narrow definition and measurement of bias. This study validated the Females In Mathematics Scale (FIMS), to measure teacher bias against females in mathematics. FIMS, based on the tripartite model of attitudes and constructed using within-method triangulation, demonstrated acceptable levels of internal reliability, and discriminant validity. Preliminary analyses showed that despite considerable variability in what teachers think and how they feel about female students in mathematics, they are more likely than not, to engage in unbiased behaviors in classroom interactions.

Keywords: Exploratory factor analysis, teacher attitudes, gender stereotypes, teacher beliefs, mathematics, tripartite model

The paucity of instruments with strong psychometric properties in educational research has raised concerns from academics about both the rigor of methods used and the validity of findings resulting from these methods for quite some time. Research on teacher attitudes, specifically bias towards students belonging to marginalized groups such as females in mathematics, and students with disabilities has been a perennial subject of discussion mainly because of mixed results yielded over several years of research. Some researchers provide evidence for the existence of teacher bias and its harmful effects on student performance (Catsambis, 1995; Inzlicht & Ben-Zeev, 2003; Sadker & Sadker, 1990), others acknowledge the presence of such bias but argue that expectancy effects resulting from bias are relatively small, accounting for 5-10% of the variance in student achievement (Jussim, 1991; Jussim & Eccles, 1992) and still, a few other studies have found no evidence to suggest that bias exists (Altermatt, Jovanovic, & Perry, 1998; Wheatley & Jones, 1990). The disparity in findings reflects potential problems in the validity and/or reliability of the measurements used, possibly due to a narrow conceptualization of attitude in the design of instruments used in the studies, and hence the limited use of triangulation in measurement.

The majority of studies on teacher bias have used dichotomous measures (Rosenfeld & Rosenfeld, 2007; Tiedman, 2000) to ascertain its existence even though research delineates attitude as a construct existing in gradations (Albarracin, Johnson & Zanna, 2005; Brophy & Good, 1974). The examination of bias through a categorical lens has a significant impact on the way bias is measured because it distills the phenomenon to its most basic form (where it is either present or absent), and as a result restricts the capabilities of what instruments can measure. We believe that 'mixed results' generated over years of research are largely a function of narrow definition and measurement of bias, and we contend that the development of robust instruments capable of capturing bias holistically hinges on how adequately bias is operationalized and how well its nature as a continuous variable is taken into account in instrument design or research methodology.

According to the Brophy and Good model (1974), teacher attitudes occur along a continuum from 'proactive' to 'over reactive' with 'proactive' teachers most inclined to being unbiased and vice versa. Good (1987) contends that rather than exist primarily at these extremes, teachers possess attitudes which vary in degrees along the continuum, the majority of who hold light expectations which are constantly adjusted as new information about the student emerges. Nevertheless, subsequent research (Altermatt, Jovanovic and Perry, 1998; Cornbleth & Korth, 1980; Hartley, 1982; Tiedman, 2000) has focused mainly on investigating the presence or absence of bias. To the best of our knowledge there have not been any studies that have taken into account the varying degrees of its existence in its measurement. Findings on bias should therefore be construed, not as 'mixed' or inconsistent, but as a testament to the nature of bias as a continuous variable and the limitations of current instruments to capture this characteristic. So far results on bias from the aforementioned studies point to individuals who lie at either end of the spectrum, and in both cases, measures have not been designed to account for the larger part of the story--one arguably central in reconciling both perspectives--the mid-section where the majority of teachers lie. Incorporating this missing piece of the puzzle in measurement requires that the operational definition of bias be expanded to include all its dimensions in measurement (i.e. its cognitive, affect and behavioral components).

Although the tripartite model defines attitudes as comprising cognitive, affective and behavioral components (Olson & Zanna, 1993), most studies have used one (Catsambis, 1995; Inzlicht & Ben-Zeev, 2003; Jones & Wheatley, 1990; Sadker & Sadker, 1990) or sometimes two (Evertson, Brophy & Good, 1973; Good & Brophy, 1972; Silberman, 1969), but not all three components in their research. A multi-dimensional approach to the operationalization and measurement of bias brings us closer to providing a more wholesome picture of the phenomenon. It also provides an avenue to use within method methodological triangulation to strengthen the validity of research findings. Triangulation uses multiple methods and measurement procedures in research in order to increase validity (Ma & Norwich, 2007). Methodological triangulation can be conducted using between methods or within methods techniques where the former involves using contrasting research methods to investigate a phenomenon e.g. using both a questionnaire and observation in a single study (Denzin, 1970). In within methods triangulation, varieties of the same method are used to investigate a research problem such as self-report questionnaires with two contrasting scales to measure a construct (Jick, 1979). By providing cognitive, affective and behavioral bases upon which attitudes can be evaluated; the tripartite model of attitudes enables the use of within methods triangulation to not only investigate teacher bias along these components but also strengthen the validity of these findings.

Here we briefly analyze methodological issues involved in the previous measurement of teacher bias and attempt to remedy these issues by developing and validating an instrument to measure teacher bias towards females in mathematics. The validation of the *Females in Mathematics Scale*, FIMS, uses the tripartite model of attitudes to expand the measured dimensions of bias, and within-method triangulation to increase validity. Teacher bias is not only limited to females in math and the sciences but also extends to males in subjects that are deemed stereotypically 'female' like reading (Hartley, 1982; Palardy, 1969), and to children from lower socio-economic backgrounds (Alexander, Entwisle & Dauber, 1993; Alvidrez & Weinstein, 1998). However, we chose to develop the Females in Mathematics Scale (FIMS), because of the abundance of already existing literature in the field on this topic, reports that teacher bias has a particularly profound impact on performance and learning outcomes of girls in stereotypically male sex-typed school subjects, such as math and science (Trouilloud et al, 2006), the chronic scarcity of females pursuing careers in these domains (National Science Foundation, 2000), and the fact that mathematics remains a critical component and determinant of entry into careers in physical science domains like engineering. Because teacher beliefs are filters that drive decisions regarding instruction (Fang, 1996) and have a significant impact on learning goals and outcomes of all students (Eccles, 1993; Trouilloud, Sarrazin, Martinek, & Guillet, 2002), it is important that it be correctly identified and reliably measured in order to design appropriate interventions or enhance existing professional development programs. Measurement and interventions could occur at any one of a number of levels from state or district, to school or grade level. Additionally, once these interventions are applied, a reliable, psychometrically sound instrument will also allow for the measurement of attitude change and evaluating the effectiveness of such programs.

Measuring Teacher Attitudes

Teacher bias has been attributed to low teacher expectations of students belonging to stigmatized groups, and the underperformance of these students has been linked to teacher behavior stemming from their expectations of these students (Brophy, 1987). These biases are based, not only on preconceived notions about student abilities but also based upon how teachers feel towards students; e.g., students that teachers are attached to and concerned about receive more teacher praise, less criticism and higher quality process questions than other students (Good & Brophy, 1972; Silberman, 1969). Studies on teacher beliefs have been conducted either through observation (Good & Brophy, 1972; Altermatt et al., 1998) or by the use of self-reports (Cook, Cameron & Tankersley, 2007; Cornbleth & Korth, 1980; Rosenfeld & Rosenfeld, 2008). In both methods comparisons between contextually stereotype-relevant groups (minorities, females, students with disabilities) and other groups have been central in determining the existence or non-existence of bias (Cornbleth & Korth, 1980; Hillman & Davenport, 1978; Jones & Wheatley, 1990). Observational studies have dwelled on differences in teacher behavior in teacher-student interactions (Altermatt et al, 1998; Evertson, Brophy & Good, 1973; Good & Brophy, 1972; Silberman, 1969) while self-report studies have relied mainly on comparisons of teacher ratings of students' ability or achievement potential to report teacher bias (Tiedmann, 2000; Alexander et al, 1993; Alvidrez & Weinstein, 1998).

Measures from Observational Studies. The results from observational studies have been mixed. Some studies indicated that teachers not only overestimated the ability of boys (Jussim & Eccles, 1992), but they also spent more quality academic time with them (Catsambis, 1995; Inzlicht & Ben-Zeev, 2003; Sadker & Sadker, 1990). Other studies found no evidence of gender bias in teacher judgments ((Hoge & Butcher, 1983; Hoge & Coladarci, 1989) or teacher-student interactions among students in the classroom (Altermatt, Jovanovic and Perry, 1998).

Altermatt et al., (1998) observed six science teachers (3 men, 3 women) across six classrooms and 165 students for 40 minutes each class over one year to investigate gender differences in both the nature of questions asked (high ordered versus low-ordered) and the frequency with which either gender was asked to respond to questions. After factoring in volunteering rates, the authors found no difference in response rates of the students who responded the most; 53% of total male volunteering compared to 48% for females, and no differences in the nature of questions asked. These findings are corroborated by similar studies, which have found no significant sex differences in opportunities students had to answer abstract questions (Hillman & Davenport, 1978; Jones &Wheatley, 1990). Results based on observational studies seem to vary, possibly because considerable subjectivity is involved in the interpretation of behavior. Second, behavior is highly contextual and there could be other factors influencing teacher behavior besides bias. Altermatt et al. (1998) proposed an alternate theory: teacher behavior may be responsive to and influenced by student behavior within the classroom, like gender differences in student volunteering rates rather than bias itself. Additionally, although observational studies provide rich accounts of the phenomenon being studied, the likelihood that the behaviors observed during these periods do not closely mimic routine behavior cannot be ruled out with certainty. This then raises concerns about ecological validity. For these reasons, considerable caution should be exercised in making generalizations about teacher bias based only on observation of teacher behavior. A more desirable alternative would be to use methodological triangulation, where observation in tandem with other methods is utilized to investigate and strengthen the validity of findings on bias.

Self-report Measures. Because research on bias has focused largely on teacher-student interactions, the use of self-reports has not been as common in research on teacher attitudes as observational studies have been. Nevertheless, self-report data have been used to explore teacher beliefs about 'weak' students (Rosenfeld & Rosenfeld, 2007), females in mathematics (Tiedmann, 2000), racial minorities (Cornbleth & Korth, 1980) and students with disabilities (Cook, Cameron & Tankersley, 2007). While the reliability of findings based solely on observational methods have been somewhat tainted by the subjectivity involved in interpreting results, the reliability of self- reports are rooted in the psychometric properties of the instruments used. Methodologists recommend that instruments be validated prior to their application in research studies (Henson, 2006) for good reason: the rigorous validation process involves item modification guided by the use of robust factor analytic techniques and validation with several hundred teachers. When conducted correctly, instrument validation allows researchers to develop reliable instruments with strong psychometric properties that enhance overall reliability and validity of findings. Without validated instruments, inferences drawn about teacher bias are tenuous, especially with small sample sizes. Unfortunately, a good number of studies have not validated instruments or failed to report sufficient information about psychometric properties of their instruments, making it difficult to evaluate reliability and validity evidence. With the exception of Cook, Cameron and Tankersley's (2007) instrument rating teacher's attitudes towards students with disabilities, the measures and reports on teacher bias have been based on non-validated instruments (Cornbleth and Korth, 1980; Rosenfeld & Rosenfeld, 2007; Tiedmann, 2000).

Cornbleth and Korth (1980) found evidence of teacher bias towards black and female students in integrated classrooms. However, detailed information on the development of the items or the theoretical basis for the attitudinal categories used to develop the 12-item instrument was not provided, making it difficult to evaluate content validity. Additionally, inferences about bias were based on evaluations of a sample of six student teachers, which is so small that any conclusions drawn from the study are likely to be non-generalizable. Similarly, Tiedmann's (2000) measure of teacher beliefs as predictors of children's concept of their mathematical ability in elementary school sampled only 28 teachers across 28 elementary school classes with a one-item measure asking teachers to evaluate students' mathematical ability on a 5-point scale. Here single responses were matched to actual math grades for 189 students and used to report higher ability perceptions for boys than for girls despite gender similarities in performance. However, this study used a small sample size in conjunction with single-item measures; this combination tends to yield non-replicable results (Fabrigar, Wegener, MacCallum & Strahan, 1999). Hence it is possible that methodological artifacts influenced the findings in these studies.

When it comes to instrument validation using factor analytic techniques, researchers suggest having at least 3-5 measured variables representing each common factor included in a study (MacCallum, 1999; Velicer & Flava, 1998) and a sample size of at least 200 for items with communalities less than .6 (MacCallum, Widaman, Zhang & Hong, 1999) or a ratio of 5-20 respondents per variable (Stevens, 1996). They recommend and the use of principal axis factoring in the first stages of instrument development because it identifies latent constructs underlying measured variables- a primary goal at this stage. A look at the literature finds little evidence of such practices: Fabrigar et al. (1999) reviewed articles published from 1991 through 1995 in two prominent journals known for methodological rigor with surprising results. About 40% of analyses did not include reports of the reliability, a large number of which were single-item measures. In numerous cases, the sample sizes used were so insufficient that they greatly increased the likelihood of obtaining under determined factors. They also found that approximately half of the published applications reported used the wrong factor analytic technique (principal components analysis vs. principal axis factoring) given their research goals. The study by Cook and colleagues (2007) faced similar challenges; they used a sample of 50 teachers to validate a four- item, four-factor scale (one item per factor) measuring teachers' attitudes towards students with disabilities. Teachers rated a total of 156 students with disabilities and 4 students without disabilities in each of their classrooms on a 4-point Likert scale with responses ranging from not at all true to extremely true for four statements: "I would like to keep this student for another year for the sheer joy of it"; "I would like to devote all my attention to this student because he or she concerns me"; "I would not be prepared to talk about this student if his or her parents dropped by for a conference", and lastly "If my class was to be reduced, I would be relieved to have this student removed." The Pearson correlations for attachment, concern, indifference and rejection ratings are quite good (values of .77, .70, .71 and .74 respectively) and compared to existing measures, a considerable level of rigor was involved in the development of this instrument. However, the authors used single item measures for each of the factors and didn't use any of the conventional tools suggested for factor analysis to determine factor structure of the items. As such, we have no way of knowing if their items reflect the constructs being measured.

It appears that the development process of instruments to measure teacher attitudes have been less rigorous than desired. Apparently measures of teacher bias have been plagued with more serious problems than previously assumed: unsatisfactory sampling procedures, relatively lax methodological procedure, instruments with either unreported or weak psychometric properties, and restricted range in measurement due to partial definition of bias. Earlier on, the use of triangulation in measurement and research to improve reliability and validity of findings was advocated, and the studies reviewed above make a clear case for that. However, given the aforementioned problems instruments in research on bias possess, it is unlikely that the use of triangulation with existing measures would improve and add to the analysis. To do so would approximate pouring new wine into old wineskins. Addressing teacher bias and remedying issues related to its measurement would require that researchers return to the drawing board and develop measures that: look at bias as a multi-dimensional construct, take into account degree with which bias exists and extend its measurement beyond dichotomy, follow stringent methodological procedure in research design, and finally, subject instruments to rigorous validation to determine and strengthen psychometric properties before use in research. This task, while daunting, is by no means impossible and certainly carries with it rewards beneficial to strengthening the overall quality of findings in research on teacher attitudes.

Developing the Females In Mathematics Scale (FIMS)

Attitude research advocates the integration of cognition in the measurement and evaluation of attitudes (Antonak & Livneh, 2000). The tripartite model of attitudes, which incorporates all three components of attitude (cognition, affect and behavior), was used as a foundation for the creation of FIMS. The cognitive component of attitude was defined as an individual's ideas, thoughts, perceptions, beliefs, opinions or mental conceptualizations of the referent (Findler, Vilchinsky & Werner, 2007); affect as the amount of positive or negative feelings one has towards the referent, and behavior as one's intent or willingness to behave in a certain manner toward the referent, or the actual behavioral response (Cook, 1992).

Method

The primary goal was to create items that would minimize response bias and capture variability in responses along the seven point Likert scale, which ranged from 'strongly disagree' to 'strongly agree'. Items corresponding to each of these three components of attitudes, specific to teaching were created and operationalized as follows:

(1) Cognition: What teachers think about females' interest, effort and ability in mathematics. High values on the seven- point scale for this factor would indicate unbiased beliefs and a positive attitude about females in math and vice versa.

(2) Affect: Teachers' feelings about females' level of engagement, ability and effort in their mathematics classes. Emotions from the pleasant-unpleasant axis of the circumplex model of emotions (Russell & Barrett, 1999) were used to generate items for affect. Teachers rated their feelings about negative levels of student engagement, effort and performance in their mathematics classes. High scores denote teachers' strong unpleasant feelings about poor performance or low engagement and effort levels of females in mathematics, indicative of high expectations and hence a more favorable attitude and vice versa.

(3) Behavior: Items that loaded under this factor concerned how teachers respond to females in their mathematics classes. Items for this subscale were based on Good's (1987) work regarding how teachers communicate expectations to students and also on the Good and Brophy (1974) model. Items evaluated differential teacher behavior in interactions between perceived high and low achievers such that high values indicated biased behaviors and vice versa.

To summarize, unbiased teachers would have very high scores on cognition and affect, and low scores on behavior.

Fifty items were created and used in a pilot study with 100 pre-service and practicing math teachers. These items made direct comparisons between males and females such as, "Females have a harder time focusing on math than males do." Participants also provided qualitative feedback on the items at the end of the survey. Both qualitative feedback and results from the factor analysis were used in tandem to assess the strength and uni-dimensionality of the items. Results signaled weaknesses in the way items were worded.

Although the factor analysis yielded factors that denoted cognition, affect and behavior, an extra factor emerged that contained a mix of cognitive and behavior items. However, the cognition items had to do with 'behavioral' indicators of females in math e.g. Females don't put as much effort in math as males and females have harder time focusing on math than males. These comparison statements, even though intended to convey a single meaning, appeared to have been interpreted multiple ways by different participants, subsequently yielding numerous multidimensional items in the factor analysis, with approximately equal loadings on one or two other factors. For example, selecting 'strongly disagree' for a statement like 'Females are more engaged in math than males' was interpreted in one of two ways: that either (a) females were not as engaged as males were (our intended meaning) or that (b) both males and females were equally engaged in math. Therefore, the meaning of both disagree and neutral were unclear. There was a general lack of variability in responses to the items. That is, individuals tended to either strongly disagree or strongly agree with items, with very few selecting more moderate response choices. Participants also indicated confusion at the implications of selecting 'neutral' for items like the aforementioned. Selecting 'neutral' for 'females are more engaged in math than males', for example, was construed in myriad ways by participants: (a) that they did not know if females were more engaged than males, (b) that they were not sure whether females were more engaged than males, or that (c) males and females were equally engaged in math. In sum, using direct comparisons between males and females led to items that were unclear, subsequently restricting the range of responses. For example, all the affect stems referenced gender differences but these questions did not account for teachers who might not perceive any gender differences in or out of their classrooms. To eliminate this ambiguity and mitigate response bias, an entirely new set of items that referenced only females was generated for this study.

Content Validity

There were 33 items (sixteen cognition, seven behavior and ten affect) went through three rounds of content reviews: a peer review, content validation by five experts, and a final review by the investigators. The content experts checked either 'not relevant', 'somewhat relevant', or 'very relevant' to gauge the relevance of each item as a measure for cognition, affect or behavior. A decision had been made a priori to make 75% the appropriate cut off for correct item placement, with the exception only for items that were believed to contribute conceptually to the factors. Any items placed in the correct category by less than 75% of the experts were deleted from the final instrument and Content Validity Indices, CVI's (McKenzie, Wood, Kotecki, Clark & Brey, 1999) were computed based only on the percentages in the 'very relevant' category. But for one affect item, all items had CVI's of 0.8 and above. Cognition and Behavior generated equally high scores while 'affect' had the lowest scores. Based on qualitative feedback from the experts and scores on the CVI, three negatively worded items in the hypothesized 'affect' factor were deleted, and replaced with positively worded items. The final number of items remained the same.

Sample and Procedures

The final survey consisted of thirty-three items: 16 dealing cognition, 7 for behavior, and 10 for affect. Items were rated on a seven point Likert scale; with responses ranging from strongly disagree to strongly agree. A convenience sample of 195 mathematics teachers teaching at the elementary to high school level, in school districts in Connecticut were contacted via email by the State Department of Education and requested to complete the FIMS online. Responses were received from a total of 195 mathematics teachers in 35 school districts representing a mix of urban (30.7%), suburban (54%) and rural schools (15.3%) of low (26.9%), middle (56%) and high (17.1%) economic status. The majority of respondents (97.2%) were from public schools. 58.2% of the sample was from High school, 24.3% from Middle school and 17.5% were from elementary School. A little more than two thirds of the sample was female (67.9%). The racial composition of the sample, which is representative of teachers in Connecticut, was predominantly White (92.7%). Other ethnicities represented included Asians (2.6%), African Americans (2.1%), Hispanic (0.5%) and 2.1% listed their ethnicity as 'other'.

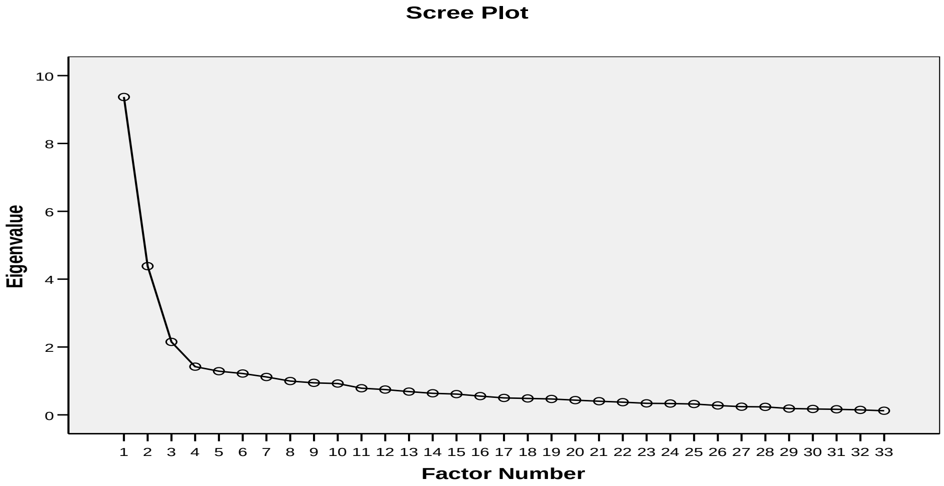
An exploratory factor analysis using principal axis factoring with oblique rotation was conducted. Principal axis factoring (PAF) was used instead of principal components analysis (PCA) because the latter is a data reduction technique that does not discriminate between unique and shared variance-- and as a result latent factors aren't the focus of the analysis (Henson, 2006). PAF was also used because it is less likely than PCA to inflate estimates of variance accounted for by the given factors (Henson, 2006). Also, the use of oblique rotation is recommended if correlations between factors are expected (Pett, Lackey & Sullivan, 2003), as was the case with factors in the FIMS.

Five different procedures under the PAF were used to determine the factor structure: Kaiser-Meyer-Olkin measure of sampling adequacy, eigenvalues greater than one, scree plot, parallel analysis (PA) and the minimum average partial (MAP) test. PA and the MAP test are not only the more accurate methods of factor extraction, but also complementary to each other: in the worst-case scenario parallel analysis tends to over extract while the MAP test tends to do just the opposite (Hayton, Allen & Scarpello, 2004). Hence the use of both methods in guiding decisions about factor extraction reduces the risk of over or under extraction.

Results

The FIMS KMO value of 0.863 suggested it was reasonable to run a factor analysis on the data and Bartlett's test of sphericity was statistically significant (p <.001). The eigenvalues above one reported seven factors and the total variance explained by all seven factors was 52.91%. The scree plot (see Fig.1), parallel analysis and MAP test, however, all suggested extracting three factors.

Figure 1
Scree plot



Three Factor Solution. A three-factor principal axis factor analysis with oblimin rotation was performed and the solution converged in 16 iterations, explaining 43.25% of the total variance. Factor 1 (cognition) explained 27.5%, factor 2 (behavior) explained 11.49% of the variance and factor 3 (affect) explained 4.71% of the total variance.

Analysis. Initially there were sixteen cognition, seven behavior and ten affect items. After the EFA was run, the pattern matrix showed that 17 items loaded on cognition, 10 on behavior and 5 on affect suggesting potential problems with cross loadings among factors. In addition to this, a close examination of the factors in the pattern matrix shown in table 1 revealed evidence of multi-dimensionality among factors, discussed in detail below.

Table 1
Pattern Matrix for the Three-Factor Solution

	Cognition	Behavior	Affect
q39. Most females are motivated in math	.88		
q30. Most females are engaged in math class	.81		
q22. Most females like taking math classes	.81		
q41. I am pleased by the level of effort most females devote to math	.80		
q34. Most females spend the time necessary to master mathematics	.76		
q28. Most females work hard at math	.74		
q37. I am delighted with the performance of most females in math	.73		

Cont. Table 1

q11. Most females find math enjoyable.	.71		
q14. Most females exert the effort required to succeed in math	.71		
q20. Most females do well in math	.70		
q27. I am happy with the level of interest most females show in math	.69		
q15. Most females are interested in math	.69		
q10. Most females are focused in math classes.	.66		
q40. Most females are capable of mastering advanced math topics with ease	.53		
q13. Understanding math concepts comes naturally for females	.53		
q9. In general most females have an easy time learning math concepts	.50		
q26. Females who excel in math are rare	.37		
q23. I tend to provide additional step by step support to females when solving simple math problems		.74	
q38. I give females additional support as they work through math problems		.72	
q16. I give females extra time to solve math problems		.58	
q31. When females struggle with math problems, I'm quick to provide them with clues to the solution		.56	
q19. I feel sorry for females who do not do well in math		.53	
q18. When correcting females in math, I try not to sound too negative		.43	
q29. I pity females who do not succeed in math		.41	
q25. I encourage females to do more math practice outside class		.37	
q33. I regret my efforts when females show no interest in math		.34	
q24. Usually, when a girl does poorly in math, it's because she does not have the ability			
q17. I am disappointed when females are bored in math class		.35	.63
q32. It is upsetting when females neglect math		.31	.58
q35. I get frustrated when my attempts to get females involved in math do not work		.33	.55
q21. I take it personally when females in my class do not meet achievement standards in math		.41	.44
q36. Usually, when a girl does poorly in math, it's because she hasn't put forth adequate effort			.41
q12. I'm very direct when providing negative feedback to females			

Note. Extraction method: Principal axis factoring. Rotation method: Oblimin with Kaiser normalization. Rotation converged in 16 iterations. Suppressed values below .30

Cognition. All but two of the original cognition items loaded under the hypothesized factor. Of the two items that failed to load on this factor, one loaded under affect (q36) and the other (q24) did not load on any factors at all. Also, three affect items cross- loaded very highly on cognition (q27 =.76, q37 =.73 and q41 = .80). A possible explanation for this is that the wording tapped into both constructs inadvertently contained a cognitive component. That is, they seemed to evaluate one's thoughts about one's affect also (e.g. I am delighted by the performance of females in math) toward females in math.

Behavior. Similarly, the behavior items appeared to be quite cohesive; but one of the original items loaded under the hypothesized behavior factor. The item that did not load on this factor (q12) also did not load on any other factors. However, seven affect items cross-loaded onto behavior (q17,q19, q21, q29. q32, q33, q35), and just as in the cognition factor, these items seemed to have wording problems. That is, the items were worded in such a way that they contained behavioral components (e.g. I feel sorry for females who do not do well in math).

Affect. The affect subscale performed less optimally than cognition and behavior subscales. Only four of ten items originally created for the scale loaded on the factor, with seven others as already discussed, cross loading on cognition and behavior. The items under affect were also multidimensional loading on both affect and behavior (see table 2).

The cognition factor emerged strongest and explained the most variance (27.5%) in teacher beliefs. Affect items performed relatively poorly compared to the rest of the other items because of the way they were worded. Affect was assessed relative to teachers' behaviors or thoughts about females in mathematics and this seems to have been the underlying cause of multi-dimensionality or cross loading in this case. Nevertheless, after excluding the cross loaded items, the pattern and structure matrices for the three-factor solution respectively, also showed moderate to high coefficients between the items and their respective factors.

Table 2
Structure Matrix for the Three-Factor Solution

	Cognition	Behavior	Affect
q39. Most females are motivated in math	.89		
q30. Most females are engaged in math class	.82		
q41. I am pleased by the level of effort most females devote to math	.80		
q22. Most females like taking math classes	.80		
q34. Most females spend the time necessary to master mathematics concepts	.76		
q37. I am delighted with the performance of most females in math	.76		
q28. Most females work hard at math	.75		
q14. Most females exert the effort required to succeed in math	.72		
q27. I am happy with the level of interest most females show in math	.71		
q11. Most females find math enjoyable.	.70		
q20. Most females do well in math	.68		
q15. Most females are interested in math	.66		
q10. Most females are focused in math classes.	.66		

Cont. Table 2

q13. Understanding math concepts comes naturally for females	.52		
q40. Most females are capable of mastering advanced math topics with ease	.52		
q9. In general most females have an easy time learning math concepts.	.50		
q26. Females who excel in math are rare	.39		
q23. I tend to provide additional step by step support to females when solving simple math problems		.72	
q38. I give females additional support as they work through math problems		.71	
q16. I give females extra time to solve math problems		.57	
q19. I feel sorry for females who do not do well in math		.55	
q31. When females struggle with math problems, I'm quick to provide them with clues to the solution		.55	
q25. I encourage females to do more math practice outside class	-.31	.44	
q29. I pity females who do not succeed in math		.44	
q18. When correcting females in math, I try not to sound too negative		.43	
q33. I regret my efforts when females show no interest in math		.39	
q24. Usually, when a girl does poorly in math, it's because she does not have the ability			
q17. I am disappointed when females are bored in math class		.38	.67
q32. It is upsetting when females neglect math		.36	.62
q35. I get frustrated when my attempts to get females involved in math do not work		.38	.60
q21. I take it personally when females in my class do not meet achievement standards in math		.47	.50
q36. Usually, when a girl does poorly in math, it's because she hasn't put forth adequate effort			.41
q12. I'm very direct when providing negative feedback to females			

Note. Extraction Method: Principal Axis Factoring. Rotation Method: Direct Oblimin with Kaiser Normalization. Suppressed values below .30

Pattern and Structure Matrices. Item-factor loadings on the pattern matrix were moderate to strong and ranged from .49 - .88, .40 -.74 and .40 -.60 for cognition, behavior and affect respectively. The moderate to strong range of values for these items relative to their factors suggests a strong relationship between the items and their respective factors after controlling for other un-related factors and variables. Similarly, the structure matrix shown in table 2 revealed moderate to strong item-factor loadings for cognition (.50 - .89), behavior (.44 - .72) and affect (.30 -.65), indicating moderate to strong correlations between items and their respective factors. There wasn't that much difference between the pattern and structure matrix with respect to the item- factor loadings, which suggests low correlations between the factors.

Factor Correlations. Table 3 shows factor correlations between cognition, affect and behavior. Cognition and behavior have a low, negative correlation because they were scored in opposite directions. That is, behavior items were scored such that high values were indicative of bias while for cognition, the reverse would be true. The relationship between cognition and affect appears to be weak and negative (implying that those who report more unbiased thoughts about females in mathematics actually harbor more biased feelings towards these females). Similarly, the positive and relatively moderate correlation between behavior and affect would suggest that those who report unbiased feelings about females in mathematics actually tend to behave in a more biased manner towards these females. There seems to be a discrepancy regarding the relationship of affect with the other two factors. This could either be a result of either social desirability in self-reporting or a methodological artifact due to poor wording of affect items, or both. The low correlations between the factors as shown in the factor correlation matrix above are, however, indicative of discriminant validity.

Table 3
Factor Correlation Matrix

Factor	Cognition	Behavior	Affect
Cognition	__	-.12	-.06
Behavior		__	.15
Affect			__

Communalities. Overall, communalities-- the variance in each item explained by its factor-- were moderate to high (Fabrigar et al., 1999) for cognition (.29 -.79.), behavior (.21 -.55) and affect (.23 -.68).

Table 4
Communalities for the three- factor solution

	Initial	Extraction
q9. In general most females have an easy time learning math concepts	.52	.29
q10. Most females are focused in math classes.	.66	.44
q11. Most females find math enjoyable	.71	.54
q12. I'm very direct when providing negative feedback to females	.37	.08
q13. Understanding math concepts comes naturally for females	.48	.31
q14. Most females exert the effort required to succeed in math	.69	.56
q15. Most females are interested in math	.66	.51
q16. I give females extra time to solve math problems	.46	.37
q17. I am disappointed when females are bored in math class	.57	.54
q18. When correcting females in math, I try not to sound too negative	.43	.21
q19. I feel sorry for females who do not do well in math	.44	.34
q20. Most females do well in math	.63	.54
q21. I take it personally when females in my class do not meet achievement standards in math	.50	.40
q22. Most females like taking math classes	.74	.66
q23. I tend to provide additional step by step support to females when solving simple math problems	.55	.55

Cont. Table 4

q24. Usually, when a girl does poorly in math, it's because she does not have the ability	.20	.03
q25. I encourage females to do more math practice outside class	.42	.32
q26. Females who excel in math are rare	.49	.33
q27. I am happy with the level of interest most females show in math	.66	.55
q28. Most females work hard at math	.75	.57
q29. I pity females who do not succeed in math	.40	.23
q30. Most females are engaged in math class	.72	.68
q31. When females struggle with math problems, I'm quick to provide them with clues to the solution	.42	.29
q32. It is upsetting when females neglect math	.47	.46
q33. I regret my efforts when females show no interest in math	.43	.23
q34. Most females spend the time necessary to master mathematics concepts	.70	.59
q35. I get frustrated when my attempts to get females involved in math do not work	.48	.48
q36. Usually, when a girl does poorly in math, it's because she hasn't put forth adequate effort	.39	.24
q37. I am delighted with the performance of most females in math	.73	.64
q38. I give females additional support as they work through math problems	.54	.51
q39. Most females are motivated in math	.80	.79
q40. Most females are capable of mastering advanced math topics with ease	.54	.31
q41. I am pleased by the level of effort most females devote to math	.78	.68

Reliability analysis. Based on the results from the factor analysis, items to conduct a reliability analysis were selected. An item was excluded from the reliability analysis if it did not load on its hypothesized factor, if it loaded on its factor but with a value less than .33 and if it had communalities less than .35 coupled with low loading on its factor. Acceptable (moderate) communality coefficients range between .4 and .7 (Fabrigar et al., 1999) so the above range was used as a baseline to retain items with communalities of .35 and above. Twenty-three items (fourteen cognition, six behavior and three affect), rated on a seven-point Likert scale were used to run the reliability analysis in SPSS. Corrected item total correlations, squared multiple correlations, average inter item correlations (IIC), standard deviations, the inter-item correlations matrix, a desirable alpha level of 0.8 set a priori, and Cronbach's alpha if item is deleted informed decisions about the reliability analysis of the subscales.

Table 5
Summary of reliability analysis for FIMS subscales

Subscale	No. Items	Alpha	CI95	Avg. IIC	SD (IIC)	Mean	SD
Cognition	14	.92	.90-.94	.46	.14	4.65	.79
Affect	3	.75	.68-.81	.51	.00	4.29	1.39
Behavior	6	.73	.66-.79	.32	.10	4.96	1.03

Cognition. This subscale reported a Cronbach's alpha of 0.92, (for more details see table 5). It had moderate to high corrected item total correlations ranging from .36 to .84, an average inter-item correlation mean of .46 and standard deviation of .14, in addition to moderate squared multiple correlations (.22 - .76). Alpha if deleted did not suggest that deleting any existing items would increase alpha significantly so all fourteen cognition items were retained.

Affect. The affect subscale consisted of three items, with a Cronbach's alpha of 0.75. The corrected item-total correlations were moderate (.57 - .60) and the same was true for squared multiple correlations that had values ranging from .33 to .36.

Behavior. The behavior subscale had a reliability coefficient of .73, and reported moderate item total correlations among its items (.33 -.62). As was the case with the affect subscale, alpha if deleted did not provide any items that would increase reliability of the scale if deleted so no items were removed.

All 23 items used to conduct the reliability analysis were retained. Based on calculations using the Spearman-Brown Prophecy formula, it was determined that the affect scale would need one additional item to bring its alpha to .8 and behavior would need to have three additional items to obtain an alpha of .8.

Discussion

So, are teachers biased towards females in mathematics? And if so, to what extent? The development of FIMS was inspired by two gaps in the literature on teacher attitudes: the general lack of psychometrically sound instruments to measure bias as multi-dimensional construct and the limited ability of current measures to report bias in gradations.

Bias. A comparison of composite scores of FIMS subscales across gender and grade levels (elementary, middle school and high school) shows that overall teachers tend to possess neutral attitudes towards female students in their classes. Items were measured on a 7-point scale with a score of 4 as neutral with values above 4 indicative of increasing favorable attitudes and vice versa. Table 6 shows comparisons of means and standard deviations by group. In general, mean scores for the subscales were within the neutral to slightly favorable range. Behavior reported slightly higher scores than cognition or affect, indicating that teachers are more likely to engage in unbiased behaviors in student interactions regardless of their beliefs or feelings.

Table 6
Composite sub-scale scores by group compared against FIMS composite score

FIMS Sub Scales	Males	Females	Elementary School	Middle School	High School	Composite Scores
Cognition:						
Mean	4.51	4.72	4.90	4.81	4.50	4.65
SD	0.71	0.82	0.85	0.83	0.71	0.79
Affect:						
Mean	4.43	4.23	4.14	4.19	4.38	4.29
SD	1.08	1.52	1.65	1.42	1.27	1.39
Behavior:						
Mean	4.77	5.06	4.90	4.83	5.11	4.96
SD	0.97	1.05	1.06	1.15	0.94	1.03

Results also show little difference in means for cognition and behavior between male and female teachers or even among teachers in elementary, middle or high school. Although mean scores for affect were generally lower than for cognition and behavior (M = 4.29, SD = 1.39 vs. M = 4.65, SD = .79 and M = 4.96, SD = 1.03, respectively), group means for affect varied more compared to the group means for cognition and behavior. A comparison of means across gender reveals higher affect means for male teachers (M = 4.43, SD = 1.08 vs. females M = 4.23, SD = 1.52). Similar comparisons across grade levels also indicate higher affect means high school teachers (M = 4.38, SD = 1.27) compared to middle school (M = 4.19, SD = 1.42) and elementary school teachers (M = 4.14, SD = 1.65). These results suggest that compared to their respective reference groups (gender and grade level) males (and high school teachers) possess slightly more positive than neutral feelings towards their students. The differences are not significant though, since all groups generally were closer to neutral than slightly favorable in their evaluative judgments, feelings and behaviors towards female students in mathematics. These results bolster the theory that teachers attitudes are neutral, and that over time most teachers adjust their expectations accordingly as more information about the student becomes available (Good, 1987).

Bivariate correlations between the subscales (see table 7) yielded non-significant correlations between cognition and affect and also cognition and behavior. However, the relationship between affect and behavior was significant. Additionally, the statistically significant affect-behavior relationship and non-significant cognition-behavior relationship seems to point to affect, rather than cognition, as a more likely factor influencing teacher behavior in interactions with students. That teachers' affect toward student engagement and effort influences teacher behavior towards students suggests that teacher behavior is contextual and is dependent on student behavior. These results support findings that highlight the role and influence of student behavior on teacher behavior (Altermatt et al., 1998), and open the door for researchers to take a closer look at the relationship between affect and behavior in the classroom.

Table 7
Subscale correlations

		Cognition	Affect	Behavior
Cognition				
	Pearson Correlation	__	-.11	.01
	Sig. (2-tailed)		.14	.94
Affect				
	Pearson Correlation		__	.31**
	Sig. (2-tailed)			.00
Behavior				
	Pearson Correlation			__
	Sig. (2-tailed)		.	

Note. ** p <.001 (2-tailed).

Although at first glance the affect-behavior relationship presented in this study seems to support previous research on these relations, it does not. This is primarily because of the difference in questions asked and hence conclusions drawn. Research prior to this study addressed how teachers felt about specific students and compared that to their behavior whereas this study examined how teachers felt about specific student behaviors. Past studies reported that students who teachers liked tended to receive more praise and higher quality questions than other students (Good & Brophy, 1972; Silberman, 1969), implying teacher exclusion of students that they lacked positive affect for, from valuable class discussions or interactions. The moderate, positive, highly significant affect-behavior relationship in this study on the other hand, demonstrates just the opposite: that the more strongly teachers feel about lower levels of engagement, effort and performance in mathematics from a group of students, the more likely they are to engage in behaviors that are empowering, communicative of high expectations toward these students, all behaviors which have been previously linked with students that teachers like (Good & Brophy, 1972) and also with males as opposed to females (Sadker & Sadker, 1989).

Degree of bias. The FIMS scale is still in its initial stages of development and as such, cannot offer individual assessment of degree of bias. It does, however, provide insight on general trends of degree by group. Analyses of means and standard deviations show that even though most teacher attitudes towards females in math generally tend to be neutral, there is considerable variability in beliefs, affect and behavior among teachers. Of notable significance here is the spread in the measures, especially for the affect subscale.

Standard deviations for the cognition subscale were smaller than those of affect or behavior, and within close range (SD = 0.71- 0.85) for males, females and different grade levels. Standard deviations for behavior were equally homogenous (SD = 0.94 - 1.15) across all groups. On the other hand, the standard deviations for affect revealed a higher level of heterogeneity among the groups sampled with values ranging from 1.08 to 1.65 compared to the other subscales. This seems to suggest that while teachers have similar beliefs about female students there is great variability with respect to how they feel about the abilities, effort and level of interest these students show in mathematics, especially among elementary school teachers (SD = 1.65) and female teachers (SD = 1.52) compared to other grade levels, and males, respectively.

Limitations and Recommendations for Future Research

The development of affect items for the validation of FIMS was a challenge. A number of affect items were multidimensional, loading on other factors. This greatly reduced the number of affect and behavior items that were used in the reliability analysis. The scale could have benefited from more measures for these constructs. Additionally, the scale has only undergone an exploratory factor analysis and is in its early stages of development. Findings presented here cannot and should not be generalized to the teacher population before confirmatory factor analyses on different samples have been conducted.

Finally, although FIMS was developed to evaluate teacher bias towards students in mathematics, it is not mathematics specific. It can and should be adapted to evaluate teacher beliefs about students belonging to groups in other disciplines for which negative stereotypes apply, after it has been formalized.

Conclusion

The American Association of University Women's latest publication, 'Why so few?' reports that despite considerable gains made by women in math and science, their success in these fields remains obstructed by stereotypes and cultural biases (AAUW, 2010). The report highlights the role of learning environments on the interest and achievement of females in math and science. That is, females tend to perform below their potential when they perceive environmental cues that highlight negative stereotypes about their ability in math, and that these effects diminish in low-stereotype contexts where gender similarities in math ability are emphasized. These results mirror empirical findings that show that situational factors like sex-composition (Inzlicht & Ben-Zeev, 2003), and features of the physical environment (Murphy, Steele & Gross, 2007) reduce females' sense of belonging to science, technology, engineering and math (STEM) domains.

The documented impact of learning environments on females' achievement and engagement in mathematics has yielded numerous efforts on the part of educators, to optimize learning for females. Currently, interventions geared towards mitigating the effects of stereotypes on female performance in mathematics e.g. teaching them about how stereotypes negatively affect performance, have been successful in diminishing these effects (Johns, Schmader & Martens, 2005). Nevertheless, the achievement of non-threatening mathematics learning environments cannot be attained through student interventions alone, especially if teacher bias is still a problem. Because teachers are a significant part of the learning environment, teacher bias can not only have a detrimental impact on student achievement and learning outcomes (Brophy, 1987), but it can also affect how female students are evaluated in these domains (AAUW, 2010). Student-focused interventions alone are not sufficient because learning environments constitute not only teachers and students but also interactions between both groups. Therefore, the effects of these student- focused interventions could further be enhanced by teacher interventions aimed at alleviating bias. Making teachers aware of their biases and giving them the tools they need to alleviate these biases will improve the learning environment for females in mathematics, bringing us closer to creating mathematics and science friendly learning environments.

However, teacher-based interventions have been scarce, partly because an evaluative measure of bias that takes into account its multi-dimensional nature has not been available. As discussed, previous attempts to measure bias ignore the multidimensional character of attitudes. This study attempted to develop an instrument that would do exactly that. The evaluation of bias as a multi-dimensional construct sheds new light on teacher attitudes towards females in mathematics. Specifically, the use of cognition, affect and behavior as evaluative bases of attitude provide valuable insight on which component is most likely to drive teacher bias, if present. Despite the limitations discussed earlier, the Females In Mathematics Scale, FIMS, is a strong first step in its endeavor to bridge the gap in empirical findings on teacher bias. As an instrument it has demonstrated relatively strong psychometric properties for the measures of all 3 dimensions of bias and attempted to establish degree by demonstrating where teachers lie on a continuum along each of these dimensions. So far FIMS has not only displayed acceptable levels of discriminant validity, but its assessment of bias has also lent support to theories by Good & Brophy (1974), Good (1987), and supported findings by Altermatt et al., (1998).

In conclusion, the FIMS has the potential to facilitate robust analysis on teacher bias in later stages of its development. It holds promise to inform both theory and practice with respect to the mechanics of attitude formation and change in regards to teacher bias. Once the validation is completed, the FIMS should be able to better pin point which dimensions of bias need to be targeted in interventions for particular groups, as important attitudinal dimensions are included and measured by the scale.

References

Albarracín, D., Johnson, B. T., & Zanna, M. P. (Eds.) (2005). *The handbook of attitudes*. Mahwah, NJ: Lawrence Erlbaum.

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1993). First grade classroom behavior: Its short and long-term consequences for school performance. *Child Development, 64,* 801-814.

Altermatt, R. E., Jovanovic, J., & Perry, M. (1998). Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. *Journal of Educational Psychology, 90*, 516-527.

Alvidrez, J. & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731-741.

American Association of University Women. (2010). *Why so few?: Women in science, technology, engineering and mathematics.* (Research Report. No. 077-10 5M). Washington, DC. Hill, C., Corbett, C., & St. Rose, A.

Antonak, R. F., & Livneh, H. (2000). Measurement of attitudes towards persons with disabilities. *Disability and Rehabilitation, 22,* 211–224.

Brophy, J. (1987). Synthesis of research on strategies for motivating students to learn. *Educational Leadership, 45*(2), 40-48.

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart & Winston.

Catsambis, S. (1995). Gender, race, ethnicity and science educations in the middle grades. *Journal of Research in Science Teaching, 32,* 243-257.

Cook, G. B., Cameron, L. D., & Tankersley, M. (2007). Inclusive teachers' attitudinal ratings of their students with disabilities. *The Journal of Special Education, 40*, 230–238.

Cook, D. (1992). Psychological impact of disability. In R. M. Parker & E. M. Szymanski (Eds.), *Rehabilitation counseling basics and beyond* (pp. 249–272). Austin, TX: PRO-ED.

Cornbleth, C., & Korth, W. (1980). Teacher perceptions and teacher-student interaction in integrated classrooms. *Journal of Experimental Education, 48,* 259-263.

Denzin, N. K. (1970). *The research act in sociology.* Chicago: Aldine.

Eccles, J. S. (1993). School and family effects on the ontogeny of children's interests, self perceptions, and activity choice. In J. Jacobs (Ed.), *Nebraska Symposium on Motivation: Vol. 40. Developmental perspectives on motivation* (pp. 145-208). Lincoln, NE: University of Nebraska Press.

Evertson, C., Brophy, J., & Good, T. (1973). *Communication of teacher expectations: Second grade.* (Report 92). The University of Texas at Austin, Austin: Research and Development Center for Teacher Education.

Fabrigar, R. L., Wegener, T.D., MacCallum, C. R., & Strahan, J. E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research, 38,* 47–65.

Findler, L., Vilchinsky, N., &Werner, S. (2007). The Multidimensional Attitudes Scale Toward Persons with Disabilities (MAS). *Rehabilitation Counseling Bulletin, 50*(3), 166-176.

Good, T. L. (1987). Two decades of research on teacher expectations: Research and future directions. *Journal of Teacher Education, 38,* 32-47.

Good, T. L., & Brophy, J. E. (1972). Behavioral expression of teacher attitudes. *Journal of Educational Psychology, 63,* 617-624.

Hartley, D. (1982). Ethnicity or sex: Teacher definitions of ability and reading comprehension in an E.P.A. primary school. *Research in Education, 28*, 9-24.

Hayton, J. C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in Exploratory Factor Analysis: A tutorial on parallel analysis. *Methodological Resources, 7,* 191-205.

Henson, R. (2006). Use of Exploratory Factor Analysis in published research. *Educational and Psychological Measurement, 66*(3), 393-416.

Hillman, S. B., & Davenport, G. G. (1978). Teacher-students interactions in desegregated schools. *Journal of Educational Psychology, 70,* 545-553.

Hoge, R., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement level. *Journal of Educational Psychology, 76,* 777-781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59,* 297-313.

Inzlicht, M., & Ben-Zeev, T. (2003). Do high achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95(4)*, 796-805.

Jick, T. D. (1979). Mixing qualitative and quantitative Methods: Triangulation in action. *Administrative Science Quarterly, 24(*4), 602-611.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, *16* (3), 175-179.

Jones, G. M., & Wheatley, J. (1990). Gender differences in teacher-student interactions in science classrooms. *Journal of Research in Science Teaching*, *27,* 861-874.

Jussim, L. & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63,* 947-961.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 93*, 54-73.

Ma, A. & Norwich, B. (2007). Triangulation and Theoretical Understanding. *International Journal of Social Research Methodology, 10* (3), 211- 226.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99.

McKenzie, J. F., Wood, L. M., Kotecki, E. J., Clark, K. J., & Brey, A. R. (1999). Establishing content validity: Using qualitative and quantitative Steps. *American Journal of Health and Behavior, 23*(4), 311-318.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science and engineering settings. *Psychological Science, 18*(10), 879- 885.

National Science Foundation. (2000). *Women, minorities, and persons with disabilities in science and engineering* (NSF Rep. No. 00-327). Arlington, VA: Bleeker, M. M.

Olson, J. M., & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44,* 117-154.

Palardy, J. M. (1969). What teachers believe- what children achieve. *Elementary School Journal, 69,* 370-374.

Pett, M. A., Lackey, R. N., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in healthcare research.* London, United Kingdom: Sage.

Rosenfeld, M., & Rosenfeld, S. (2008). Developing effective teacher beliefs about learners: the role of sensitizing teachers to individual learning differences. *Educational Psychology, 28*, 245-272.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76,* 805-819.

Sadker, M., & Sadker, D. (1990). Confronting sexism in the college classroom. In S. K. Gabriel & I. Smithson (Eds.), *Gender in the classroom: power and pedagogy* (pp. 176-187). Urbana: University of Illinois Press.

Silberman, M. L. (1969). Behavioral expression of teachers' attitudes toward elementary school students. *Journal of Educational Psychology, 60,* 402-407.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (*3^{d} ed.*). Mahwah, NJ: Lawrence Erlbaum.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*(1*),*144-151.

Trouilloud, D., Sarrazin, P., Bressoux, P., & Bois, J. (2006). Relation between teachers' early expectations and students' later perceived competence in physical education classes: Autonomy supportive climate as a moderator. *Journal of Educational Psychology, 98*(1), 75-86.

Trouilloud, D., Sarrazin, P., Martinek, T. G., & Guillet, E. (2002). The influence of teacher expectations on students' achievement in physical education classes: Pygmalion Revisited. *European Journal of Social Psychology, 32*(5), 591-607.

Velicer, W. F., & Flava, J. L. (1998). Effects of variable and subject sampling as factor pattern recovery. *Psychological Methods, 3,* 231-251.

About the Authors

Katherine Picho is a doctoral candidate in the Cognition and Instruction program at the University of Connecticut. Her research interests focus on Stereotype Threat, and gender differences in leadership style. Katherine's areas of interest in quantitative research methods include hierarchical linear modeling, structural equation modeling, meta-analysis and psychometrics.
Email: edpsychresearch@gmail.com

Dr. Katrichis is an associate professor in Marketing in the Barney school of business at the University of Hartford.
Email: Katrichis@hartford.edu

Dr. D. Betsy McCoach is an associate professor in the Measurement, Evaluation and Assessment program at the University of Connecticut. She has extensive experience in hierarchical linear modeling, instrument design, factor analysis, and structural equation modeling. Betsy co-edited Multilevel Modeling of Educational Data. She has published numerous journal articles and book chapters in the field of education. Betsy is the current co-editor of the Journal of Advanced Academics. She also serves on the editorial review boards for the American Educational Research Journal, the Journal of Educational Psychology, the Journal of Educational Research, and Gifted Child Quarterly.
Email: betsy.mccoach@uconn.edu