Factorial, Convergent, and Discriminant Validity of TIMSS Math and Science Motivation Measures: A Comparison of Arab and Anglo-Saxon Countries

Article *in* Journal of Educational Psychology · February 2013

DOI: 10.1037/a0029907

CITATIONS
80

READS
1,415

15 authors, including:

Herb Marsh
Australian Catholic University
618 PUBLICATIONS **67,014** CITATIONS
SEE PROFILE

Adel Abduljabbar
King Saud University
43 PUBLICATIONS **1,142** CITATIONS
SEE PROFILE

Maher M Abu-Hilal
Sultan Qaboos University
74 PUBLICATIONS **561** CITATIONS
SEE PROFILE

Alexandre J S Morin
Concordia University Montreal
211 PUBLICATIONS **6,551** CITATIONS
SEE PROFILE

Journal of Educational Psychology
2013, Vol. 105, No. 1, 108–128

0022-0663/13/$12.00 DOI: 10.1037/a0029907

Factorial, Convergent, and Discriminant Validity of TIMSS Math and Science Motivation Measures: A Comparison of Arab and Anglo-Saxon Countries

Herbert W. Marsh
University of Western Sydney, King Saud University, and University of Oxford

Adel Salah Abduljabbar
King Saud University

Maher M. Abu-Hilal
Sultan Qaboos University

Alexandre J. S. Morin
University of Western Sydney

Faisal Abdelfattah
King Saud University

Kim Chau Leung
Hong Kong Institute of Education

Man K. Xu
University of Cambridge

Benjamin Nagengast
University of Tübingen

Philip Parker
University of Western Sydney

For the international Trends in International Mathematics and Science Study (TIMSS2007) math and science motivation scales (self-concept, positive affect, and value), we evaluated the psychometric properties (factor structure, method effects, gender differences, and convergent and discriminant validity) in 4 Arab-speaking countries (Saudi Arabia, Jordan, Oman, and Egypt) and 4 English-speaking Anglo-Saxon countries (United States, England, Australia, and Scotland). In this article, we also highlight methodological weaknesses in the TIMSS approach to these motivation measures. We found reasonable support for within-group invariance across the math and science domains and between-group invariance across countries for full factor loading invariance and partial item intercept invariance. However, the factor structure is complicated by strong negative-item method effects and correlated unique characteristics associated with the use of math and science items with parallel wording. Correlations involving the motivation factors were reasonably similar across countries, supporting both discriminant and convergent validity in relation to achievement, plans to take more coursework in math and science, and long-term educational aspirations. However, gender differences largely favor girls in the Arab countries (with strong single-sex education systems) relative to Anglo countries (and international norms). The juxtapositions of latent mean differences in achievement and motivation factors are perplexing; students from Anglo countries had substantially higher achievement than students from Arab countries but had substantially lower motivation across all 8 math and science factors.

Keywords: math and science motivation, trends in international mathematics and science study, math and science gender difference, negative item method effects, cross-cultural measurement invariance

This article was published Online First September 17, 2012.

Herbert W. Marsh, Centre for Positive Psychology and Education, University of Western Sydney, Kingswood, New South Wales, Australia; Psychology Department, Education College, King Saud University, Riyadh, Saudi Arabia; Department of Education, University of Oxford, Oxford, United Kingdom. Adel Salah Abduljabbar, Department of Psychology, College of Education, King Saud University. Maher M. Abu-Hilal, College of Education, Sultan Qaboos University, Al Khoudh, Oman. Alexandre J. S. Morin, Centre for Positive Psychology and Education, University of Western Sydney. Faisal Abdelfattah, The Excellence Research Center of Science and Mathematics Education and Department of Psychology, College of Education, King Saud University. Kim Chau Leung, Department of Special Education and Counselling, Hong Kong Institute of Education, Tai po, New Territories, Hong Kong. Man K. Xu, Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom. Benjamin Nagengast, Center for Educational Science and Psychology, Department of Education, University of Tübingen, Tübingen, Germany. Philip Parker, Centre for Positive Psychology and Education, University of Western Sydney.

The authors would like to thank Matthias Von Davier, Anna Preuschoff, Michael Martin, Geert Hofstede, Misho Minkov, and Patrick Alexander for helpful comments at earlier stages of this research.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Centre for Positive Psychology and Education, University of Western Sydney, Bankstown Campus, Locked Bag 1797, Penrith South DC, New South Wales 1797, Australia. E-mail: h.marsh@UWS.edu.au

Supplemental materials: http://dx.doi.org/10.1037/a0029907.supp

The present investigation is a substantive-methodological synergy (Marsh & Hau, 2007), bringing to bear current methodology in order to address substantively important issues in relation to the math and science motivation constructs (self-concept, affect, and value) of students from four Arab countries (Saudi Arabia, Oman, Jordan, and Egypt) and how they compare with responses by students from four English-speaking Anglo countries (U.S., Australia, England, and Scotland), using the international Trends in International Mathematics and Science Study (TIMSS2007) data. Substantively, this study draws on theory and research in self-concept and expectancy-value theory. Methodologically, we critique and improve upon the methodology used by TIMSS to create scores for these constructs that are widely used in secondary data analyses. Because the TIMSS data have a strong cross-cultural perspective, it is also important to test the cross-cultural generalizability of theoretical models (e.g., Marsh & Hau, 2003, 2004; Marsh, Hau, Artelt, Baumert, & Peschar, 2006; Seaton, Marsh, & Craven, 2009; van de Vijver & Leung, 1997, 2000).

The focus on Arab countries is particularly relevant to researchers from these countries, as there has not previously been a psychometrically rigorous evaluation of TIMSS motivation constructs for Arab countries. Furthermore, the juxtaposition with Anglo countries is of broad interest to cross-cultural researchers. In particular, these motivation constructs and achievement have been juxtaposed in many studies, particularly comparing the United States with Japan, China, and other East Asian countries and have demonstrated what has been referred to as paradoxical findings (e.g., Shen & Tam, 2006; also see Minkov, 2008; Stevenson, Chen, & Lee, 1993; Stevenson & Stigler, 1992). More specifically, even though academic self-concept is systematically related to academic achievement within each country, students from Anglo countries have consistently higher self-concept but lower levels of academic achievement. However, there has not been much research juxtaposing motivational constructs and achievement based on comparison of Anglo and Arab countries, using nationally representative samples responding to appropriately matched materials (but see Abu-Hilal & Bahri, 2000). A particularly salient feature of the educational systems in Arab countries is that students are taught in single-sex schools in which male teachers teach boys and female teachers teach girls.

The Importance of Motivational Self-Beliefs in Education

As noted by Marsh, Craven, Hinkley, and Debus (2003) and many others (e.g., Bøe, Henriksen, Lyons, & Schreiner, 2011; Bong, 1996; Heyman & Dweck, 1992), there is no consensus as to what are the key motivational constructs, what they are called, and how they are to be measured. Marsh et al. (2003) described the *jingle-jangle fallacy*, in which two scales with the same label might not refer to the same construct and two scales with different labels might refer to the same construct. Similarly, Bong (1996) suggested that in order to avoid what she referred to as a "conceptual mess" (p. 152), researchers should apply confirmatory factor analysis (CFA) and structural equation models to evaluate the structural, predictive, convergent, and discriminant validity of different motivation constructs. Although they have previously been labeled in inconsistent ways, motivational constructs in the TIMSS survey can be broadly divided into *self-concept* (sometimes referred to as *competence beliefs*), *positive affect* (sometimes referred to as *intrinsic motivation*), and *task value* (a construct combining value and importance but overlapping substantially with what some refer to as *extrinsic motivation*). A main focus of our study is to evaluate the measurement invariance of these motivational constructs. However, we are particularly interested in self-concept, its specificity in relation to math and science domains, how it is related to and distinguished from other motivational constructs, and the structure of these motivational constructs. Self-concept is widely accepted as an important universal aspect of being human and as central to understanding the quality of human existence (Bandura, 2006; Bruner, 1996; Marsh & Craven, 2006; Pajares, 1996; Pajares & Schunk, 2005). Positive self-beliefs are valued as a desirable outcome in many disciplines of psychology and are central in the positive psychology revolution sweeping psychology, which focuses on how individuals can optimize their life (Diener, 2000; Fredrickson, 2001; Seligman & Csikszentmihalyi, 2000). In assessing the construct (convergent and discriminant) validity of these constructs, we focus particularly on criteria measured separately for math and science (achievement and plans to pursue further study) but also on other correlates such as gender and long-term educational aspirations.

Domain Specificity and Discriminant Validity

Within the academic self-concept literature, there is strong support for the domain specificity of self-concept, an important component of motivation—particularly the separation of math and verbal self-concepts, as demonstrated with the Organization for Economic Cooperation and Development's Program for International Student Assessment (OECD–PISA) research (Marsh et al., 2006; also see Marsh & Hau, 2004). However, more contestable is the implicit assumption in the design of the Self-Description Questionnaires (SDQ; Marsh, 2007) that is the basis of much of this research: that affect and competency components of each domain together form a single, unidimensional factor. Thus, the math, verbal, and school self-concept scales for the SDQ were measured by three sets of parallel items designed to reflect competency and affective components. Competency was defined by items asking students whether they were good at, learned things quickly in, got good marks in, and found work to be easy in different school subjects. Affect was defined by items asking students whether they were interested in, looked forward to, liked, and enjoyed work in different school subjects. Marsh, Craven, and Debus (1999; also see Arens, Yeung, Craven, & Hasselhorn, 2011) reviewed a number of studies supporting some separation of the affect and competency components of self-concept. They found that two large cohorts of students aged 7–13 years showed that correlations between verbal and math self-concepts systematically decreased with age (i.e., became more domain specific) but that the relation between affect and com-

petencies within each domain remained very high ($r = .75$). On the basis of this research, they concluded, "We tentatively recommend that researchers distinguish between competency and affect components of academic self-concept, qualified by the need to evaluate further the construct validity of this separation in relation to additional external validity criteria" (p. 567).

Other research on motivation constructs (e.g., Deci & Ryan, 1985; Eccles, 1983; Eccles & Wigfield, 2002; Feather, 1982; Renninger, 2000, 2009; Renninger, Hidi, & Krapp, 1992; Stipek & Mac Iver, 1989) provides a clear theoretical rationale for the separation of competency self-beliefs from affective components of motivation such as intrinsic motivation, interest, and task value. Motivation research based on expectancy-value theory is particularly relevant to the present investigation. Although Eccles et al. (1983) originally hypothesized expectations of success and self-concept to be separate constructs, subsequent research has indicated that the two constructs could not be separated empirically (Eccles & Wigfield, 1995; Wigfield & Eccles, 2002). Within this expectancy-value framework, Eccles, Wigfield, and colleagues (Eccles & Wigfield, 1995; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Wigfield, 1994; Wigfield & Eccles, 2000; Wigfield et al., 1997) showed that correlations between self-perceived competency and interest were evident even for very young children but that the size of this relation increased with age during early school years. In expectancy-value theory (e.g., Wigfield & Eccles, 1992; Wigfield, Eccles, & Pintrich, 1996), task value is differentiated into different components: intrinsic value (interest and enjoyment), utility value (perceived future usefulness), attainment value (importance of the activity and of succeeding at it), and cost (negative consequences such as fear of failure or lost opportunities). Both expectancy and value constructs are highly domain specific (Eccles et al., 1993). Researchers have found that although the domain specificity of different constructs tended to increase as individuals age, the relations between expectancy and value remained high or even increased with age (e.g., Eccles & Wigfield, 2002; Eccles et al., 1993; Wigfield & Eccles, 2002; Wigfield et al., 1996; Wigfield, Tonks, & Klauda (2009). While self-perceived competence was related to several different value constructs in the expectancy-value model, the relations with intrinsic value were consistently strongest (Wigfield & Eccles, 2002).

Whereas expectancy beliefs have been shown to be closely associated with performance in both cross-sectional and longitudinal studies, interest and value beliefs are sometimes better predictors of choice, effort, and persistence at achievement-related activities (e.g., Bøe et al., 2011; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Meece, Wigfield, & Eccles, 1990; Nagengast et al., 2011; Parsons, Adler, & Meece, 1984; Trautwein & Lüdtke, 2007; Wigfield et al., 1997). This separation between competency and intrinsic motivation is also central to Deci and Ryan's (1985, 1991) cognitive evaluation theory. From their theoretical perspective, they posited that "we should expect a close relationship between perceived competence and intrinsic motivation such that the more competent a person perceives him- or herself to be at some activity, the more intrinsically motivated the person will be" (Deci & Ryan, 1991, p. 58).

A Cross-Cultural Perspective: Generalizability to Arab Countries

The generalizability of particularly Western self-concept research findings to Arab countries has been the focus of previous research by Abu-Hilal and colleagues (Abu-Hilal, 2001; Abu-Hilal & Aal-Hussain, 1997; Abu-Hilal & Bahri, 2000). Abu-Hilal and Bahri (2000) evaluated the generalizability of responses to Marsh's SDQ instrument for responses by elementary and junior high school students from United Arab Emirates (UAE). They found support for the a priori factor structure but noted that the self-concept factors were less differentiated (more correlated) than typically is found in Western research. For example, math and verbal self-concept scales on the SDQ are typically almost uncorrelated (Marsh, 2007). In contrast, even for the older junior high school sample, these two academic scales were moderately correlated ($r = .37$) in UAE, but less so than the substantial correlation between corresponding measures of achievement ($r = .62$). Nevertheless, there was good support for the convergent and discriminant validity of the math and verbal self-concept scales in relation to corresponding measures of math and verbal achievement.

Abu-Hilal and Bahri (2000) have suggested the idea (based in part on Sharabi, 1975) that Arab students—particularly boys—are socialized in the home and school in a way that "does not seem to encourage students to be independent: it does not give children the opportunity to evaluate themselves" (p. 319). Anecdotally, the authors noted that when asked to evaluate their skills and performances in different school subjects, several students commented, "Are you sure you want us to judge our performance? I think that teachers can tell you better than we can" (p. 320). Thus, compared with Western students of a similar age, these Arab students were less knowledgeable about their relative strengths and weaknesses. More specifically, self-concepts were more uniformly high (i.e., less differentiated in terms of level within each domain) and more highly correlated across the different domains of self-concept. Abu-Hilal and Bahri (2000) indicated that this pattern of results is similar to that found with younger children in Western research (e.g., Marsh, 1989, 1990), in which self-concepts of young children are also uniformly high and substantially correlated but become more differentiated with age as children obtain more experience relative to their related strengths and weaknesses. Commenting on gender differences in achievement, Abu-Hilal (2001) noted that during adolescence, Arab boys have much more freedom than girls and that this lack of freedom meant that girls focus more on schoolwork than do boys (also see Hassan & Khailifa, 1999). Consistently with these expectations, girls had substantially higher scores for both verbal and math achievement tests but also reported putting more effort into schoolwork, were more motivated in school, and had slightly higher math and verbal self-concepts. Nevertheless, boys tended to be praised more than girls, and this contributed to their high self-concepts, even though girls exerted more academic effort than boys. Abu-Hilal (2001) suggested that this pattern of results and cultural influences generalizes reasonably well across other Middle Eastern Arab countries.

TIMSS2007 (Olson, Martin, & Mullis, 2008): Background to the Present Investigation

Factor Structure and Motivation Scale Construction

Since its inception in 1959, the International Association for the Evaluation of Educational Achievement (IEA) has conducted a series of international comparative studies designed to provide policy makers, educators, researchers, and practitioners with information about educational achievement and learning contexts. TIMSS2007 is the fourth in a cycle of internationally comparative assessments dedicated to improving teaching and learning in mathematics and science for students around the world. Carried out every 4 years at the fourth and eighth grades, TIMSS provides data about trends in mathematics and science achievement over time. In each cycle, the primary focus has been on substantive, theoretical, and methodological excellence in achievement tests—appropriate tests in the two subject domains that provide a basis for comparing achievement across countries and over time.

Each cycle has also included a student questionnaire that measures student motivation in mathematics and science—the focus of the present investigation. Ideally, the construction of multidimensional instruments should be based on theory, item and reliability analysis, exploratory and confirmatory factor analyses, tests of convergent and divergent validity, validation in relation to external criteria, and application in research and practice. From a construct validation perspective, theory, measurement, statistical analysis, empirical research, and practice are inexorably intertwined, so that the neglect of one will undermine the others (see Marsh, 2007). Unfortunately, the high standards for the achievement tests developed with TIMSS are not reflected in the measures of student motivation, as there is little evidence of a strong theoretical basis for the selection and construction of items. In each successive cycle, the nature of the items has changed and psychometric evidence in support of the scores used to summarize responses to these items has been incomplete, inadequate, or simply not presented.

Liu and Meng (2010) examined the factor structure of the TIMSS2003 motivation items based on responses by eighth-grade students in three East Asian countries and the United States. Noting that motivation in both mathematics and science were represented by 12 parallel-worded items, they were critical of researchers who used these items without appropriate recognition of the underlying constructs represented by the items. In contrast to the highly sophisticated analyses of the structure underlying achievement test responses, TIMSS has not been entirely consistent in the way these items are scored in different data collections, and the *TIMSS2003 Technical Report* (Ramirez & Arora, 2004) did not present even preliminary factor analyses to justify the factor structure underlying these 12 motivation items. In response to this limitation, Liu and Meng conducted exploratory factor analyses of responses to the 12 mathematics motivation items in each of the four countries and found reasonably consistent support for a two-factor solution in which the factors were labeled *mathematics self-concept* and *perceptions of mathematics importance*. Curiously, the authors did not pursue confirmatory factor analysis methods that would have allowed them to test more formally the proposed factor structure and the invariance of factor loadings and item intercepts needed to justify the comparison of mean differences between the countries on these constructs. Nevertheless, the results do highlight the need for more rigorous tests of constructs provided in the TIMSS database (Olson, 2008), which are the basis of policy and many secondary data analyses conducted by researchers from all over the world.

For TIMSS2007, there was a slightly revised set of items (see Table 1) and a new classification of items representing a hypothesized three-factor solution for each achievement domain. One of the 12 items ("wanting to take more coursework") was not included in any of the three factors. Preuschoff, Martin, and Mullis (personal communication, October 20, 2011) noted that preliminary analyses indicated that this item was conceptually different and did not fit with any of the other factors. For present purposes, rather than simply excluding it, we have retained this item but posited it to represent a separate factor (see Table 1). The *TIMSS2007 Technical Report* (Olson et al., 2008) briefly summarized results for a confirmatory factor analysis in support of the hypothesized three-factor solution in separate analyses of responses to the math and science motivation items. The CFAs of the motivation items were a useful contribution for TIMSS2007. However, the results were only a small component of a larger study, so that the results were not presented in sufficient detail to evaluate the appropriateness of the analysis. Although current best practice is to provide a variety of goodness-of-fit indices, the authors presented only the root-mean-square error of approximation (RMSEA). Furthermore, in relation to current standards of a good fit, the RMSEA of .087 for mathematics did not reflect a good fit. In addition, the reported degrees of freedom were too low—27 for math and 26 for science, rather than 41—and not even the same for the two subject domain models, suggesting that additional unreported parameters were included in the model. Also, the correlation between affect and self-confidence (two components that were combined in at least one of the reported classifications for TIMSS2003) was sufficiently high (math $r = .724$; science $r = .883$) as to possibly detract from the discriminant validity of the two constructs. Indeed, the authors presented no clear evidence of the discriminant validity either of different constructs within the same subject domain or the same construct across the different subject domains—based on CFA models combining both math and science motivation factors. Furthermore, no multigroup analyses were reported to test the invariance of factor loadings (weak measurement invariance) and item intercepts (strong measurement invariance) over countries that would justify the comparison of means across countries, like those that are a focus of the results of achievement test items for TIMSS. Similarly, no analyses were reported to test the invariance of factor loadings and intercepts over math and science subject domains that would justify the comparison of means for the parallel math and science constructs. Hence, an important contribution of the present investigation is to present apparently the first empirical tests of the invariance of the a priori factor structure across different countries (between-group invariance) and across the math and science domains (within-country invariance).

In 2007 and earlier data collections, TIMSS used what they called a *scale method* for deriving scores for the multi-item motivation factors, which were made available to researchers for secondary data analysis and have been used in international reports. Starting with a simple average response (after reverse scoring negatively worded items), a trichomization was then applied so

Table 1
A Priori Factor Structure Relating TIMSS Motivation Items to Latent Factors

Item	Factor loading	Item wording
Mathematics		
Value		
MVAL1	0.626	I think learning mathematics will help me in my daily life.
MVAL2	0.611	I need mathematics to learn other school subjects.
MVAL3	0.743	I need to do well in mathematics to get into the university of my choice.
MVAL4	0.742	I need to do well in mathematics to get the job I want.
Self-concept		
MSCP1	0.750	I usually do well in mathematics.
MSCP2	0.857	I learn things quickly in mathematics.
MSCN1	0.550	Mathematics is harder for me than for many of my classmates.
MSCN2	0.724	I am just not good at mathematics.
Affect		
MAFFP1	0.894	I enjoy learning mathematics.
MAFFP2	0.917	I like mathematics.
MAFFN1	0.686	Mathematics is boring.
Coursework		
MMORE	1.000	I would like to do more mathematics in school (single item).
Science		
Value		
SVAL1	0.626	I think learning science will help me in my daily life.
SVAL2	0.611	I need sciences to learn other school subjects.
SVAL3	0.743	I need to do well in sciences to get into the university of my choice.
SVAL4	0.742	I need to do well in sciences to get the job I want.
Self-concept		
SSCP1	0.750	I usually do well in science.
SSCP2	0.857	I learn things quickly in science.
SSCN1	0.550	Science is harder for me than for many of my classmates.
SSCN2	0.724	I am just not good at science.
Affect		
SAFFP1	0.894	I enjoy learning science.
SAFFP2	0.917	I like science.
SAFFN1	0.686	Science is boring.
Coursework		
SMORE	1.000	I would like to do more science in school (single item).

Note. This factor analysis is discussed in greater detail in the Results section. Briefly, these results are based on Model 9 (see subsequent discussion of Model 9 in Table 3) and are average results over five imputed data sets. Factor loadings are unstandardized estimates in a model identified by constraining all factor variances to be 1.0. Factor loadings were constrained to be equal across all eight countries and constrained to be equal across the parallel-worded items for the math and science constructs. TIMSS = Trends in International Mathematics and Science Study; M = mathematics; VAL = value; SCP = self-concept (positive); SCN = self-concept (negative); AFFP = affect (positive); AFFN = affect (negative); MORE = more coursework; S = science.

that each student received one of three score values: high, medium, or low (Ramirez & Arora, 2004). Although this approach might be heuristic for some limited reporting purposes, this trichotomization of a reasonably continuous measurement scale is generally unacceptable in relation to current best practice, as it substantially reduces reliability, statistical power, and predictive validity (MacCallum, Zhang, Preacher, & Rucker, 2002). Preuschoff, Martin and Mullis (personal communication, October 20, 2011) indicated that "we understand that categorizing responses to background scales into three categories does result in loss of information, but we do it because it aids interpretability" and that for TIMSS2011, there will be scale scores based on Rasch modeling. However, this problem also undermines support for reliability presented in the technical report (trichotomized scores are substantially less reliable than scale scores) for the limited evidence for the validity of the math and science motivation scales in the *TIMSS2007 Technical Manual* and also for the secondary analyses of the public data based on these trichotomized scores (e.g., Chiu, 2011).

Method Effects: Matching Items and Negative Item Effects

Method effects are nontrait effects associated with idiosyncratic aspects of particular items or methods of data collection. Failure to incorporate method effects appropriately is likely to have substantial effects on goodness of fit, biased parameter estimates, and substantive interpretations. In the present investigation, we evaluate two possible sources of method effects associated with the use of parallel worded items to infer math and science motivation constructs and the use of a mixture of positively and negatively worded items within the same construct.

The use of negatively worded items constitutes a potential source of construct-irrelevant variance that detracts from the construct validity of interpretations of self-concept responses, particularly by young students. However, even for responses by older students and adults, factor analyses of psychological rating scales comprising a mixture of positively and negatively worded items typically reveal apparently distinct factors reflecting the positive and the negative items, respectively (see Carmines & Zeller, 1979; Chiu, 2008, 2011; Marsh, 1986, 1996). A typical approach based on the logic of multitrait-method studies of method effects is to test for negative-item method effects by including correlated uniquenesses between negatively worded items (e.g., Benson & Hocevar, 1985; Carmines & Zeller, 1979; DiStefano & Motl, 2006; Marsh, 1986, 1996; Marsh, Scalas, & Nagengast, 2010). Consistent with these expectations, Chiu (2008, 2011) reported this type of negative-item method effect for the two negatively worded self-concept items for the TIMSS2003 data. Chiu included correlated uniquenesses to control for this negative-item bias, but further noted that the factor loadings for these items were systematically lower than for the positively worded items and suggested that "items that are negatively worded appear to be unreliable in cross-cultural studies" (p. 251). In the present investigation, we compared models with and without correlated uniquenesses to test for negative item effects and corrected for them if they were shown to exist (see Figure 1; also see the online supplemental material for further documentation of correlated uniquenesses [CUs] considered and Mplus syntax).

As the wordings of the math and science motivation items in TIMSS2007 are parallel, correlated uniquenesses were also posited a priori between each matched pair of items, as recommended by Marsh and Hau (1996). Indeed, when the same item is used for multiple domains, a correlation between the unique components of

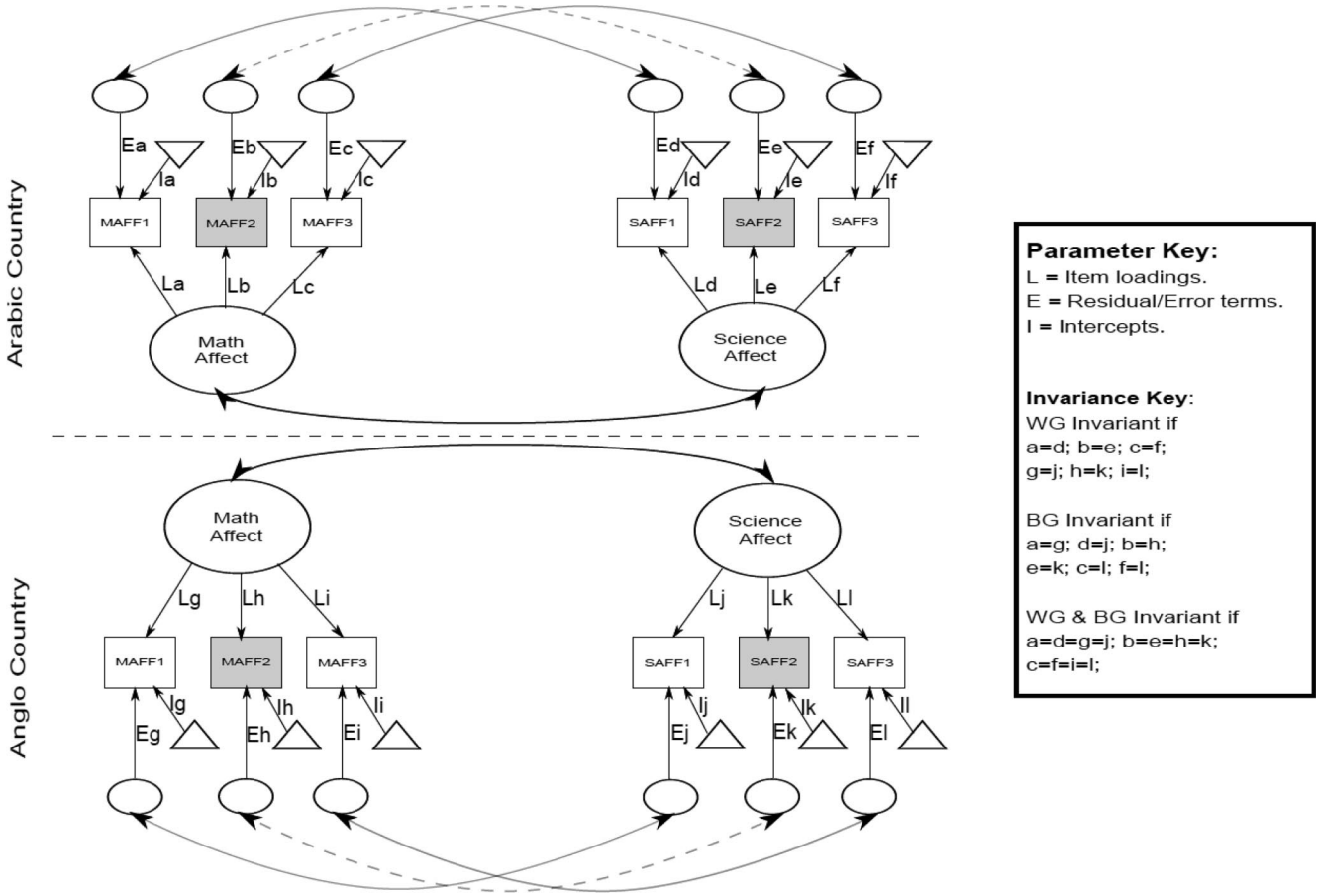

Figure 1. Conceptual schematic diagram of the statistical model for two factors—math affect (MAFF) and science affect (SAFF)—measured with parallel worded items. Large ovals = latent factors; squares = measured variables; triangles = measured-variable intercepts; small circles = measured variable uniqueness. The light gray double-headed arrows are correlated uniquenesses (CUs). The solid gray lines are CUs for math and science items with parallel wording. The dashed gray lines are CUs relating negatively worded items (shaded in gray) to control for the negative-item effect. In some cases, as in Figure 1, the two types of CUs coincide (i.e., negatively worded parallel items), but in the full model (see Appendix in the online supplementary material). CUs were used to relate all of the negatively worded items, across domains as well as within domain. A detailed set of invariance tests (see parameter key) were used to evaluate invariance of factor loadings and intercepts. Within-group (WG) tests of invariance constrained parameters to be invariant over domain, while between-group (BG) tests constrained parameters to be invariant across the eight (four Arab and four Anglo) countries.

each item on the two domains that cannot be explained by the correlations between the factors is likely to exist. The failure to include these CUs is likely to systematically bias parameter estimates such that correlations among matching latent factors across different domains are systematically inflated (Marsh, Martin, & Debus, 2001; Marsh et al., 2011; Marsh, Parada, & Ayotte, 2004). Although this is clearly a complication and a source of idiosyncratic bias in the present investigation, it is important to emphasize that there are also important advantages to the use of parallel-worded items, so long as this strategy can be used to represent the different domains appropriately. Indeed, the entire rationale for tests of invariance over domains and the appropriate tests of latent means for parallel math and science constructs depends fundamentally on the use of parallel wording (see Figure 1).

In summary, an important methodological focus of the present investigation was to address important limitations in the psychometric evaluation of math and science motivation factors in TIMSS2007 based on responses by students from Arab and Anglo-Saxon countries.

The Present Investigation: A Priori Predictions and Research Questions

In the present study, we pursued seven related psychometric and substantive research questions that underpin support for construct validity.

1. Reliability

Given that the motivation constructs come largely from Western research and empirical evidence based on previous TIMSS research, we hypothesized the coefficient alpha estimates of reliabil-

ity to be systematically lower in the four Arab-speaking countries than in the English-speaking Anglo-Saxon countries.

2. Method Effects

Following from earlier discussions, we hypothesized that there would be substantial method effects associated with the use of parallel wording in the math and science motivation items. We also hypothesized that there would be method effects associated with negatively worded items. We hypothesized that these two sets of method effects would be reasonably independent, such that controlling for both sets of method effects would improve goodness of fit more than controlling for either one separately. More generally, we hypothesized that achieving an acceptable goodness of fit would require us to control for both sets of method effects.

3. Factor Structure

We hypothesized that responses to the 24 motivation items would support an a priori factor structure of eight latent factors (see Table 1); three multi-item motivation factors (self-concept, value, and positive affect), and one single-item factor (desire to pursue further coursework) for both math and science domains. This a priori factor structure followed from our earlier discussion and the design of the TIMSS scales. However, we also expected the correlation between self-concept and affect to be so high as to possibly detract from the discriminant validity of these constructs (e.g., Marsh et al., 1999).

4. Measurement Invariance

Apparently, there has not been previous research on the measurement invariance of the TIMSS motivational constructs. This is surprising, since measurement invariance is an important component of construct validation and a prerequisite to any variance–covariance (including correlation and predictive paths) and mean-level comparisons across subpopulations (i.e., gender or countries) or domains (i.e., math vs. science). Hence, we leave as open research questions whether there is support for the invariance of factor loadings (weak invariance) and item intercepts (strong invariance) in relation to country, domain (math vs. science) and gender, and whether the relative support for invariance differs across these groupings. As illustrated in Figure 1, we differentiated between within-group tests of invariance across the math and science domains and between-group invariance based on invariance across the eight (four Arab and four Anglo) countries.

5. Country-Level Differences in Achievement and Motivation

Based on the TIMSS2007 results (Mullis, Martin, & Foy, 2008a, 2008b), we know that students from Arab countries perform more poorly on the math and science achievement tests. We leave as a research question how these lower levels of achievement translate into motivational constructs. Although it might be expected that poorer achievement should lead to lower values on the motivation constructs, the process of forming self-beliefs is complex and is not a simple function of achievement levels. Indeed, one of the perplexing findings in international comparisons of the juxtaposition of self-concept and achievement is that achievement in the United States is only moderate (particularly in relation to many Asian countries), while U.S. self-concepts are above international averages (e.g., Shen, 2002; Shen & Tam, 2006; Stevenson, Chen, & Lee, 1993; Wilkins, 2004). Part of the explanation is that there are strong frame of reference effects such that particularly academic self-concept responses are based on comparisons with other students in the same school or class (Marsh, 2007), so that country-level differences in achievement are unlikely to be reflected in individual student self-concepts. This is further complicated in Arab countries in that educational settings are single-sex, so that the frames of reference are specific to each gender (i.e., boys have little opportunity to compare their performances with girls, or vice versa). Furthermore, Abu-Hilal's (2001) research in Arab countries has suggested that the socialization process in the school and family leads Arab students—particularly boys—to be less critical of themselves, so that they have higher self-concepts than might be expected in terms of their objective achievements.

6. Gender Differences

For the international TIMSS2007 data, averaged across participating countries, there were very small gender differences in favor of women for both math and science achievement (Mullis et al., 2008a, 2008b), consistent with evidence that historically observed gender differences favoring boys in math and science achievement are declining, disappearing, or reversing in direction. Thus, we expected gender differences based on latent achievement factors to be small, particularly in the Anglo countries. Nevertheless, of particular relevance to the present investigation is the comparison of gender differences in the largely coeducational Anglo countries versus the single-sex education systems in Arab countries. Consistent with findings and speculations by Abu-Hilal (2001), TIMSS2007 results showed that almost all countries with the largest gender differences in favor of girls were Arab countries (Mullis et al., 2008a, 2008b). However, we leave as a research question whether these gender differences in achievement are also reflected in other motivational variables and their correlates (plans to pursue math and science study; long-term educational aspirations).

7. Convergent and Discriminant Validity of Motivation Factors

In support of the convergent and discriminant validity of the TIMSS motivational constructs, we hypothesized (a) academic achievement to be more highly correlated to self-concept than to value and affect, (b) desire to pursue more coursework to be more highly correlated with positive affect than self-concept, and (c) domain specificity to be demonstrated: that is, math motivation would be more highly related to math criteria than to science criteria, and science motivation factors would be more related to science criteria than to math criteria (e.g., Eccles et al., 1983; Eccles & Wigfield, 2002; Marsh, 2007; Wigfield et al., 1997, 2009).

Method

Participants

TIMSS2007 (Olson et al., 2008) assessed the competencies in mathematics and science for nationally representative samples of

students from 59 participating countries (for more details about the processes underlying the development of the TIMSS2007 instruments, translation of materials, sampling, data collection, scaling, and data analysis, see the *TIMSS 2007 Technical Report* by Olson et al., 2008). The basic sampling design is a two-stage cluster design consisting of sampling of schools and sampling of intact classrooms from the target grade in the school. For present purposes, participants were eighth-grade students from Saudi Arabia (4,269 students, 47% male, from 203 intact classrooms), Jordan (5,251 students, 47% male, from 199 intact classrooms), Oman (4,752 students, 53% male, from 157 intact classrooms), Egypt (6,582 students, 51% male, from 237 intact classrooms), the United States (7,593 students, 50% male, from 509 intact classrooms), England (4,048 students, 48% male, from 441 intact classrooms), Australia (4,103 students, 55% male, from 327 intact classrooms), and Scotland (4,205 students, 49% male, from intact 257 classrooms).

TIMSS (Olson, Martin, & Mullis, 2008) used item response theory (IRT) to scale student achievement scores in mathematics and science. In science, the content domains were biology, chemistry, physics, and earth sciences. In mathematics, the content domains were algebra, data and chance, number, and geometry. In both subject domains, slightly more than half of the achievement items (51% in mathematics and 55% in science) involved a constructed response, whereas the remaining items were multiple choice. The final items were selected on the basis of item analyses for responses from large-scale pilot studies. As noted earlier, two sets of 12 motivation items each (see Table 1) were used to measure math and science motivation constructs. Students responded to all motivation items using a 4-point (*agree–disagree*) Likert response scale. On the basis of these data, we pursued analyses to address the seven research questions discussed earlier.

Data Analysis

Weighting and clustering. All analyses were based on TIMSS's HOUWGT weighting variable, which was provided as part of the TIMSS database. HOUWGT incorporates six components: three have to do with sampling of the school, class and student, and adjustment factors associated with nonparticipation at the level of the school, class, and student. It is based on the actual number of students in each country that is appropriate for correct computation of standard errors and tests of statistical significance. In addition, the TIMSS user's guide also notes that it is appropriate to apply a correction for clustering inherent in the two-stage clustering sample. For present purposes, the eight countries were treated as grouping variables that were the basis of the multigroup analyses, whereas the class identification variable was treated as a clustering variable to control for the clustered sample (using the complex design option and robust maximum likelihood options in Mplus). Class rather than school was used as the clustering variable, because class was the sampling unit used in the TIMSS sampling design, which was based on sampling all students within intact classes. In fact, most schools were represented by a single class, and a given class might not be representative of the school from which it came.

Plausible and missing data. In the TIMSS2007 database, the achievement tests for each student are reported as five plausible values—numbers drawn randomly from the distribution of scores that could be reasonably assigned to each student. Following TIMSS protocols, we performed all data analyses with achievement separately for each of the five plausible values, and the results were aggregated appropriately in order to obtain unbiased estimates. Although the amount of missing data was relatively small, there is increasing awareness of the limitations of traditional approaches to missing data such as mean substitution, listwise deletion, or pairwise deletion for missing data. Following the logic of the plausible values, we used multiple imputations (Graham, 2009; Schafer & Graham, 2002) to deal with missing responses to the motivation items. In the imputation model, we included all variables used in any of the analyses as well as school-average measures of math and science achievement and country indicator variables. Analyses were weighted, using the TIMSS HOUWGT weighting variable. Five imputed data sets were constructed, and one of the five sets of plausible achievement scores was used with each of the imputed data sets. This strategy allowed us to consider simultaneously the appropriate implementation of the plausible values and appropriate handling of missing data in the same set of analyses. In each case, reported results are based on an appropriate aggregation of results across the multiple datasets to obtain appropriate parameter estimates, standard errors, and goodness-of-fit statistics. We note, however, that this strategy was used primarily to incorporate the multiple plausible values, as the amount of missing data was so small (an average of less than 2% for the motivation items).

Tests of factorial and measurement invariance. Comparison of results across different countries or across different domains (i.e., math and science) requires strong assumptions about the invariance of the factor structure across the groups or domains. If the underlying factors are fundamentally different, then there is no basis for interpreting observed differences (the "apples and oranges" problem). For example, in cross-national studies of motivational differences like those considered here, interpretation of mean differences—or even relations among different constructs—presupposes that the factors are the same across countries (i.e., Arab and Anglo countries). In the present investigation, we initially consider a 2×8 classification of invariance tests (see Figure 1), the invariance over the two domains (math and science), invariance over the eight countries, and invariance over all 16 (8×2) combinations of domain and country. We subsequently add gender to the evaluation of invariance over all 32 ($8 \times 2 \times 2$) combinations of domain, gender, and country.

Recent TIMSS studies have evaluated the similarity of factor loadings for motivational factors across the math and science domains. However, in these preliminary analyses, formal tests of the invariance of factor loadings were not applied. Even more surprisingly, preliminary CFAs presented in the *TIMSS2007 Technical Manual* (Olson et al., 2008) did not consider the invariance of factor loadings across multiple groups (nor, to our knowledge, has this issue been pursued elsewhere). More difficult are issues related to the invariance of item intercepts needed to justify the comparison of latent means. Although issues of noninvariance of item intercepts and differential item functioning are well known in relation to TIMSS achievement tests, these issues have been largely ignored in relation to TIMSS motivation constructs.

Following Meredith (1993) and others, Marsh (Marsh, Lüdtke, Muthén, et al., 2010; Marsh, Muthén, et al., 2009) operationalized a taxonomy of partially nested models that begins with a model with no invariance of any estimated parameters or *configural invariance;* only

the parameters fixed to zero or 1 that define the structure of the model are invariant. The initial focus is on the invariance of the factor loadings—sometimes referred to as *weak measurement invariance* or *pattern invariance*—which requires that factor loadings be invariant over groups or over time. *Strong measurement invariance* requires that the indicator intercepts and factor loadings are invariant over groups (or domains) and justifies comparison of latent means. *Strict measurement invariance* requires invariance of item uniquenesses (in addition to invariant factor loadings and intercepts) and justifies the comparison of manifest means over groups or time. Strict measurement invariance is required in order to compare manifest scale scores (or factor scores), in that differences in reliability for the multiple groups would distort mean differences on the observed scores. However, for comparisons based on latent constructs that are corrected for measurement error, the valid comparison of latent means only requires support for strong measurement invariance and not the additional assumption of invariance of measurement error. Hence, comparison of group mean differences based on latent-variable models like those considered here makes fewer assumptions than those based on manifest scores. Even less demanding is the comparison of latent correlations in different groups, a condition that only requires weak measurement invariance (i.e., invariance of the factor loadings). Although these tests require full invariance of all parameter estimates within each category (e.g., all factor loadings), Byrne, Shavelson, and Muthén (1989) argued for the usefulness of a less demanding test of partial invariance in which some parameter estimates are not constrained to be invariant.

Goodness of fit. It is now broadly accepted that all a priori models are false and will be shown to be false when tested with a sufficiently large sample size. For this and other reasons, chi-square tests of statistical significance are of little relevance for evaluation of goodness of fit for a single model and are even more problematic for the comparison of fit for two different models that require additional assumptions that are unlikely to be met (e.g., Marsh, Balla, & McDonald, 1988). Hence, in applied CFA and structural equation modeling research, there is a predominant focus on indices that are sample size independent (e.g., Marsh, Balla, & Hau, 1996; Marsh et al., 1988; Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004) such as the RMSEA, the Tucker–Lewis index (TLI), and the comparative fit index (CFI)—as well as the chi-square test statistic and an evaluation of parameter estimates. The TLI and CFI vary along a 0-to-1 continuum and values greater than .90 and .95 typically reflect acceptable and excellent fit to the data, respectively. RMSEA values of less than .05 and .08 reflect a close fit and a minimally acceptable fit to the data, respectively. However, for purposes of model comparison, comparison of the relative fit of models imposing more or fewer invariance constraints is more important than the absolute level of fit for any one model—so long as the fit of the best fitting model is acceptable. Cheung and Rensvold (2001, 2002) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, then there is reasonable support for the more parsimonious model. Chen (2007; Chen, Curran, Bollen, Kirby, & Paxton, 2008) suggested that when the RMSEA increases by less than .015, there is support for the more constrained model. For indices that incorporate a penalty for lack of parsimony, it is also possible for a more restrictive model to result in a better fit than a less restrictive model. Although we relied on these guidelines in the present investigation, it is important to emphasize that these are only rough guidelines (Marsh, Hau, & Wen, 2004), and it is recommended that applied researchers use an eclectic approach based on a subjective integration of a variety of different indices—including the chi-square, detailed evaluations of the actual parameter estimates in relation to theory, a priori predictions, common sense, and comparison of viable alternative models specifically designed to evaluate goodness-of-fit in relation to key issues. This is consistent with the approach we used here.

Extended models. In the extended models considered here, we further evaluated the construct validity of TIMSS motivation factors by relating the latent motivation factors to participant background variables (gender and long-term educational aspirations) and math and science achievement test scores (see Olson et al., 2008). In addition, we considered as validity criteria the two variables asking students whether they wanted to pursue further coursework in mathematics and science (see Table 1). For present purposes, we adapted a multiple-indicator-multiple-indicator-cause (MIMIC) approach (see Jöreskog & Sörbom, 1993; Kaplan, 2000) in which each of six correlates (gender, educational aspirations, math and science achievement, math and science coursework plans) was related to the motivation latent factors in order to test a priori hypotheses (see Research Question 7). The CFA MIMIC approach is like the multiple regression approach, but it is stronger in that it is based on latent constructs that are purged of measurement error and controlled for method bias (as well as providing a test of the underlying model rather than just assuming that it is correct in the formation of scale scores). Variables added to this model include gender (1 = female, 2 = male), long-term educational aspirations (EDASP), math achievement (MACH: a composite of algebra, data and chance, number, and geometry), and science achievement (SACH: a composite of chemistry, earth science, biology, and physics).

Estimation. All analyses in the present investigation were conducted with Mplus Version 6.1 (Muthén & Muthén, 2008–2011). Analyses consisted of CFAs based on the Mplus robust maximum likelihood estimator (MLR), with standard errors and tests of fit that were robust in relation to nonnormality of observations and the use of categorical variables where there were at least four or more response categories, particularly when nonnormality was not excessive and a design-based correction (Mplus's complex design option) was used to control for the nonindependence of observations (Muthén & Muthén, 2008–2011). On this basis, and because of the complexity of the models under consideration, we chose to use MLR estimation rather than a categorical estimation procedure that treats Likert responses as categorical variables rather than continuous variables. Our decision was based on the growing body of work (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Muthén & Kaplan, 1985; Rhemtulla, Brosseau-Liard, & Savalei, 2012) suggesting that categorical estimation procedures make little or no difference to the parameter point estimates that are the major focus of our study, and—particularly for very large models, as in the present case (with as many as 1,500 parameter estimates in some models)—sometimes detract from estimation precision. Further, they frequently create problems associated with convergence and proper solutions in testing for the invariance of thresholds. Also, skews and kurtoses in our study were not extreme (e.g., Hau & Marsh, 2004); across the 24 motivation items, the average skew was −0.76 (with none of the skews being more than 1.7) and the average kurtosis was −0.22 (with only one item having a kurtosis greater

than 2 in absolute value). Nevertheless, we acknowledge the use of MLR estimation as a potential limitation of the study.

Results

Reliability of TIMSS Math and Science Motivation Scales

In response to our first research question, we began by evaluating the reliability of the TIMSS2007 motivation scales in the two groups of countries. Particularly given that the motivation constructs come largely from Western research, it would not be surprising to find the reliability estimates higher in Anglo countries than Arab countries. Due in part to the brevity of the six multi-item motivation scales, each consisting of only three or four items, at least some of the coefficient alpha (α) estimates of reliability (Table 2) are not acceptable. Reliabilities for scales reached a desirable standard of .80, but in other cases fell below an acceptable value of .70. In general, the reliabilities were lower in Arab countries (mean α across the six scales in four countries = .648) and higher in the Anglo countries (mean α = .831). In each case, the reliability in the international sample consisting of all participating countries (mean α = .75) reported in *TIMSS2007 Technical Manual* (Olson et al., 2008) fell between the values for the Arab and Anglo samples considered here. Of particular worry are the unacceptably low alphas for Arab students responding to self-concept items (mean α = .553). This suggests, perhaps, problems in the definition of these constructs. It is interesting that these two scales were quite reliable for the Anglo countries (mean α = .816) but were also relatively low, based on the international sample across all participating countries (αs = .73 and .66) as reported in *TIMSS2007 Technical Manual* (Olson et al., 2008). The low levels of reliability for Arab countries (and for the international sample more generally) are worrisome, particularly for analyses based on manifest scores. Furthermore, the trichotomized scale scores actually used in TIMSS reports, and apparently the basis of many secondary data analyses, are likely to be substantially less reliable. These differences in reliability between Arab and Anglo countries—but also the differences in reliability for the different motivation constructs—would undermine the validity of interpretations based on manifest scale scores. From this perspective, it is critical that such comparisons be based on latent-variable models that appropriately control for unreliability—like those considered here.

Table 2
Reliability of TIMSS Math and Science Motivation Constructs Used in This Study

	Mathematics			Science			
Country	Self	Affect	Value	Self	Affect	Value	Mean
Four Arab countries							
Saudi Arabia	.521	.739	.696	.488	.717	.717	.646
Jordan	.664	.774	.692	.634	.764	.764	.715
Oman	.507	.676	.708	.485	.608	.608	.599
Egypt	.546	.697	.620	.581	.677	.677	.633
Mean	.560	.722	.679	.547	.691	.691	.648
Four Anglo countries							
United States	.839	.855	.725	.822	.859	.859	.827
Australia	.818	.855	.784	.803	.881	.881	.837
England	.800	.865	.718	.841	.885	.885	.832
Scotland	.778	.857	.744	.829	.877	.877	.827
Mean	.809	.858	.743	.824	.875	.875	.831
International	.73	.81	.70	.66	.78	.78	

Note. For the corresponding math and science scales, the wording of the items was strictly parallel (see Table 1). Reliability estimates are Cronbach's alpha estimates based on the present investigation. International value is the median coefficient alpha across all participating countries reported in *TIMSS 2007 Technical Report* (Olson, Martin, & Mullis, 2008). TIMSS = Trends in International Mathematics and Science Study.

Factor Structure of TIMSS Math and Science Motivation Scales

Our a priori model (Table 1) posits that the 24 motivation items (12 math items and 12 science items with parallel wording) can be explained by four math factors and four science factors. The four factors for each domain are the three multi-item scales in the *TIMSS Technical Manual* (self-concept, positive affect, and value) and an additional single-item factor for each domain based on the two items that were excluded by analyses in the *TIMSS Technical Manual* (wanting to pursue more study in math and in science).

In our a priori model, we posited two sources of method effects based on responses to matching items and responses to negatively worded items (see earlier discussion). In order to fully evaluate different aspects of this a priori model, we tested several models leading up to this model and then subsequent models, testing the invariance of the factor structure over the eight Arab and Anglo countries (see Table 3).

Noninvariant factor structures (configural invariance). In the present investigation, we tested a wide variety of invariance models in relation to Research Questions 2 and 3, but began with simple multigroup models that do not impose any invariance constraints. In Model 1 (M1 in Table 3), we posited that all 12 math motivation items would reflect a single math factor and that all 12 science items would reflect a single science motivation factor. As noted by Liu and Meng (2010), this is consistent with how TIMSS motivation factors have been characterized (inappropriately) in some research. However, the fit of this model was highly unacceptable, leading to its rejection. In Model 2, we posited two factors for each subject domain, which is like the structure posited by Liu and Meng (2010). However, the fit of this model was similar to that in Model 1. In Model 3, we posited four motivation factors for each domain, as posited in our a priori model. Although the fit was substantially improved, it was still unacceptable in relation to current standards. This is worrisome, in that a typical precondition for pursuing tests of invariance is good support for configural invariance. However, this result is consistent with our a priori prediction that there would be substantial method effects (Research Question 2) associated with negatively worded items and parallel-worded items.

Method effects. In Models 4–6, we pursued issues outlined in Research Question 2. More specifically, we added correlated uniquenesses to represent method effects associated with matching items (M4), negatively worded items (M5), or both (M6). The correlated uniqueness terms included in each model are summarized in Appendix 1 in the online supplemental materials. Consis-

Table 3
Summary of Goodness of Fit Statistics for Total Group Models

Model	Chi	Scale	*df*	CFI	TLI	RMSEA	Description
24 Motivation items only							
No Inv							
M1	72267	1.729	2008	.681	.650	.083	M1: 1 math & 1 science factors
M2	54454	1.706	1968	.762	.733	.072	M2: 2 math & 2 science factors (like TIMSS 2003)
M3	27322	1.684	1808	.884	.859	.053	M3: math & 4 science factors
M4	17442	1.667	1720	.929	.908	.042	M3 + CUs for parallel items
M5	13621	1.696	1688	.946	.929	.037	M3 + CUs for negative items
M6	8779	1.676	1624	.968	.956	.029	M3 + CUs for parallel & negative items
WG & BG FL Inv							
M7	9044	1.683	1688	.966	.956	.029	M6 with WG domain FL Inv
M8	10761	1.703	1736	.959	.948	.032	M6 with BG Country FL Inv
M9	11041	1.703	1747	.958	.947	.032	M6 with WG & BG
WG & BG Intercept Inv							
M10	12361	1.701	1819	.952	.942	.034	M9 + WG Inv
M11	17788	1.702	1859	.928	.914	.041	M9 + BG Inv
M12	18565	1.703	1871	.924	.911	.042	M9 + WG & BG Inv
M13	12719	1.704	1832	.952	.940	.034	M9 + WG & BG (part)[a] Inv
Motivation items and additional correlates							
Gender Inv 8 × 2 group							
M14	15619	1.645	3662	.947	.936	.036	M13 with gender: No Inv
M15	15721	1.645	3667	.947	.936	.036	M14 with gender: Inv FL
M16	16028	1.645	3697	.946	.936	.036	M14 with gender: Inv for FL & Intercept
Additional validity criteria							
M17	17731	1.667	2349	.941	.924	.036	M13 with additional correlates

Note. All analyses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within schools. Additional correlates were gender, math and verbal achievement, parent education, and educational aspirations. Inv = invariance; FL = factor loading; Chi = chi-square; *df* = degrees of freedom ratio; Scale = Mplus scaling factor used in chi-square tests of statistical significance; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation; Inv = invariance; CUs = a priori correlated uniquenesses based on the negatively worded items; WG = within group; BG = between group; FL= factor loading; TIMSS = Trends in International Mathematics and Science Study.
[a] In a test of partial invariance of intercepts over country, four invariance constraints were relaxed (SCn1, SCn2, Val1, Val3), noting that the corresponding math and science items were constrained to be equal.

tent with a priori hypotheses, comparison of the fit indices for these models indicated that each source of method effect contributed substantially and independently to the goodness of fit. Model 6, which included both sources of method effects, provided a reasonable fit to the data, supporting our a priori predictions about the factor structure (Research Question 2) and the need to incorporate method effects (Research Question 3). With the inclusion of these a priori method effects, Model 6 provides support for configural invariance and is the basis of subsequent models.

Weak measurement invariance (factor loading invariance). In Models 1–6, we did not impose any invariance constraints. In subsequent models considered in this section, we tested the invariance of factor loadings over the math and science domains (a within-group test of invariance) or country (a between group test of invariance) or both (see Figure 1). In Model 7, we tested the invariance of factor loadings over the math and science domains such that factor loadings for math items on the math factors were the same as those for the science items (with parallel wording) on the science factors. We refer to this as a within-group test of invariance because it only imposed invariance constraints within each of the country groups. This model is more parsimonious than Model 6 in that there are fewer estimated parameters—a single factor loading is estimated for each matched pair of math and science items, rather than two separate factor loadings. The fit for Model 7 was very good, and the difference in fit statistics was clearly less than the cutoff values typically used to support the more parsimonious model (Table 3). On this basis, we concluded that the factor loadings were fully invariant over domain for each country.

In Model 8, we tested the invariance of factor loadings over the eight countries (but did not require factor loadings to be invariant over subject domain). The goodness of fit was still acceptable; the decline in fit relative to Model 6 was not substantial. Thus, there was modest support for the invariance of factor loadings over country. In Model 9, factor loadings were constrained to be invariant over both domain and country. Not surprisingly, the combination of both sets of constraints resulted in a slightly poorer fit, but one that was still acceptable in relation to typical criteria of an acceptable fit and one that did not differ substantially from those in which only within or between-group constraints were imposed. Factor loadings based on this final model were presented earlier in Table 1 (also see Appendix 1 in supplemental materials for a full set of all parameter estimates). Of course, in the unstandardized metric, the patterns of factor loadings are all identical for parallel math and science items within each country (within-group domain invariance) and for items across the eight (between-group country

invariance). Reflecting the negative item effect, the negatively worded items have systematically smaller factor loadings (see related findings for TIMSS2003 data in Chiu, 2011).

Strong measurement invariance (item intercept invariance). Strong measurement invariance requires that intercepts are invariant and is an important assumption in justifying the comparison of latent means. In Models 10–13 (Table 3), we tested the invariance of item intercepts over domain (M10), over country (M11), or over both domain and country (M12). There is reasonable support of the invariance of intercepts over domain (M10), but not over country (M11) or over the combination of domain and country (M12). On the basis of modification indexes for Model 12, we explored partial invariance models for country intercepts, while assuming full intercept invariance over domain. This resulted in freeing invariance constraints across countries for four items (two of four self-concept items, two of four value items; see footnote, Table 3). Model 13, with partial invariance of intercepts, provided a reasonable fit to the data, a fit that did not differ much from Model 9, with no intercept invariance constraints. Because the metric of the latent means is established by the items with invariant intercepts, the results suggest that it is reasonable to compare mean differences between countries for these factors (as well as comparing math and science scores for these factors within each country). Nevertheless, comparisons of latent means particularly over country should be made cautiously.

Measurement invariance in relation to gender. Although the invariance over gender was not a central focus of the present investigation, the evaluation of gender differences is predicated on appropriate levels of invariance. In order to evaluate this assumption, we divided each of the eight country samples into separate subsamples for boys and girls (i.e., 16 groups = 8 countries × 2 genders). Starting with Model 13 (already considered), we began with a model with no invariance constraints over gender (Model 14), and then we imposed additional constraints to evaluate the invariance of factor loadings (Model 15) and the invariance intercepts (Model 16) over gender. The results provided very strong support for the full invariance of factor loadings and intercepts over gender (i.e., configural, weak, and strong invariance).

Construct Validity of TIMSS Math and Science Motivation Scales: Relations to Correlates

There has been surprisingly little rigorous evaluation of the convergent and discriminant validity of the TIMSS math and science motivation factors. In pursuit of this goal, we systematically evaluate support for a priori hypotheses (Research Questions 5–7) for relations between the TIMSS motivation factors with validity criteria and other correlates. Starting with Model 13 (discussed earlier), we added correlates to the model to evaluate these predictions (Table 4) based on this extended model. We begin with an evaluation of latent mean differences among the constructs in relation to two domains (math vs. science) and eight countries (with particular emphasis on the four Arab and four Anglo countries) in response to issues raised in Research Question 5. We then move to an evaluation of correlations among the six multi-item motivation factors and correlates included in the extended model in each country (see Table 4) in order to evaluate gender differences (Research Question 6) and support for the convergent and discriminant validity of responses to the TIMSS2007 motivation factors (Research Question 7).

Mean differences over domain and country for motivation constructs (Research Question 5). For purposes of evaluating latent mean differences in the four motivational constructs, we took advantage of the fact that there is parallel wording for the corresponding math and verbal constructs and good support for strong measurement invariance over domain. This allowed us to make comparisons of latent mean differences over domain as well as comparisons of country-level differences in the latent means. For purposes of these analyses, the latent means for the four math motivation constructs were fixed at zero in the Saudi sample. Thus, size and direction of differences in all the remaining 15 sets of latent mean differences (science in Saudi Arabia, math and science in the remaining seven countries) were evaluated in relation to the math constructs for the Saudi sample.

To illustrate the logic of these comparisons, we initially discuss in detail the latent means for math self-concept, which is fixed at zero in the Saudi sample but has a standardized latent mean value of –.263 in the U.S. sample. Thus, math self-concepts are significantly lower in the United States than in Saudi Arabia, and because these are standardized mean differences (*SD*s of latent constructs are 1.0 across all countries), the difference between the two countries is in standard deviation units. The corresponding values for science self-concept are .172 (Saudi Arabia) and –.335 (United States). Thus, science self-concept in Saudi Arabia is significantly higher than math self-concept in Saudi Arabia, but science self-concept in the United States is significantly lower. For all four Arab countries, self-concept, positive affect, value, and plans to pursue further coursework for both science and math are almost all significantly positive, indicating that they are as high or higher than the math motivation constructs in Saudi Arabia. In marked contrast, the eight motivation constructs are negative in all four Anglo countries, indicating that they are substantially lower than the math motivation constructs in Saudi Arabia.

Results based on Model 17 also provide a basis for comparing achievement test scores in the eight countries. Because the achievement scores were standardized (mean = 0, *SD* = 1), positive scores reflect better than average achievement across the eight countries and negative scores reflect poorer than average achievement. Consistent with a priori predictions (Research Question 5), mean achievement scores in the Anglo countries are significantly positive for both math and science (from +.256 to +.656 *SD*s above the mean) and significantly negative in almost all four Arab countries (–.938 to +.146). Indeed, only science achievement in Jordan is not significantly below the overall mean, but even this score is below the mean science achievement in each of the four Anglo countries.

Gender differences (Research Question 6). The final Model 17 allowed us to evaluate gender differences. Particularly because of the single sex nature of the Arab school systems (see earlier discussion), gender differences are of special interest. Gender differences in Table 4 are represented as correlations between gender and each of the remaining constructs (labeled *male* in Table 4, as positive correlations indicate that boys have higher scores).

Across the four Anglo countries, gender differences are mostly small to moderate, but consistently favor boys, who tend to score as high or higher than girls for all math and science constructs (Table 4). For the eight motivation constructs across four Anglo

Table 4
Estimated Latent Means and Correlations: Construct Validity (Model 17, Table 3)

Country/variable	Estimated correlation matrix for the latent variables										
	MVAL	MSC	MAFF	SVAL	SSC	SAFF	MASP	SASP	MACH	SACH	EDASP
Saudi Arabia											
Latent means											
USTD	0.000	0.000	0.000	0.264	0.172	0.122	0.000	0.044	−0.938	−0.539	0.306
SE	—	—	—	.022	.023	.033	—	.024	.024	.026	.030
STD	0.000	0.000	0.000	0.299	0.321	0.137	0.000	0.051	−1.622	−0.809	0.327
Latent correlations											
MVAL	1.000										
MSC	.556	1.000									
MAFF	.634	.772	1.000								
SVAL	.595	.276	.356	1.000							
SSC	.388	.477	.309	.622	1.000						
SAFF	.446	.278	.405	.701	.829	1.000					
Motivation validity criteria											
MASP	.521	.395	.673	.305	.141	.311	1.000				
SASP	.400	.190	.320	.584	.518	.696	.364	1.000			
MACH	.130	.516	.111	.053	.387	.088	−.009	.024	1.000		
SACH	.113	.382	.050	.103	.498	.166	−.042	.077	.574	1.000	
EDASP	.188	.285	.133	.176	.267	.121	.088	.110	.237	.258	1.000
Gender differences											
Male	.105	−.042	.094	.094	−.041	.127	.104	.128	−.142	−.266	−.164
SE	.025	.034	.033	.028	.037	.031	.028	.026	.037	.034	.025
Jordan											
Latent means											
USTD	0.301	0.043	0.369	0.507	0.119	0.314	0.388	0.275	−.173	0.146	0.253
SE	.040	.036	.045	.041	.040	.051	.033	.039	.037	.038	.033
STD	0.391	0.061	0.449	0.664	0.179	0.380	0.551	0.371	−0.228	0.180	0.280
Latent correlations											
MVAL	1.000										
MSC	.494	1.000									
MAFF	.539	.781	1.000								
SVAL	.657	.346	.321	1.000							
SSC	.406	.459	.304	.673	1.000						
SAFF	.417	.267	.357	.679	.820	1.000					
Motivation validity criteria											
MASP	.496	.364	.498	.347	.244	.266	1.000				
SASP	.395	.201	.261	.606	.570	.647	.351	1.000			
MACH	.242	.571	.277	.188	.332	.064	.145	.083	1.000		
SACH	.253	.475	.190	.243	.423	.143	.150	.145	.751	1.000	
EDASP	.293	.393	.216	.253	.330	.114	.134	.146	.443	.438	1.000
Gender differences											
Male	−.052	.020	−.024	−.084	−.042	−.066	−.042	−.082	−.098	−.170	−.059
SE	.023	.027	.028	.023	.025	.023	.021	.021	.044	.043	.031
Oman											
Latent means											
USTD	0.412	−0.008	0.548	0.683	0.057	0.488	0.452	0.384	−0.581	−0.352	−0.035
SE	.037	.028	.038	.035	.033	.039	.030	.031	.031	.036	.031
STD	0.571	−0.016	0.860	1.057	0.116	0.815	0.721	0.662	−0.814	−0.439	−0.034
Latent correlations											
MVAL	1.000										
MSC	.481	1.000									
MAFF	.564	.792	1.000								
SVAL	.664	.307	.342	1.000							
SSC	.416	.420	.307	.635	1.000						
SAFF	.432	.281	.375	.635	.860	1.000					
Motivation validity criteria											
MASP	.429	.358	.524	.316	.201	.284	1.000				
SASP	.377	.181	.265	.517	.487	.626	.354	1.000			
MACH	.337	.491	.268	.238	.336	.205	.213	.217	1.000		
SACH	.342	.411	.200	.265	.421	.209	.216	.182	.714	1.000	
EDASP	.175	.229	.140	.127	.237	.107	.111	.101	.299	.322	1.000
Gender differences											
Male	−.153	.003	−.095	−.101	.081	−.017	−.088	−.071	−.281	−.322	−.132
SE	.030	.031	.031	.029	.041	.034	.023	.024	.033	.033	.026

Table 4 (*continued*)

Country/variable	Estimated correlation matrix for the latent variables										
	MVAL	MSC	MAFF	SVAL	SSC	SAFF	MASP	SASP	MACH	SACH	EDASP
Egypt											
Latent means											
USTD	0.073	0.149	0.543	0.375	0.214	0.482	0.379	0.311	−0.453	−0.471	−0.024
SE	.034	.031	.038	.033	.030	.037	.031	.030	.029	.034	.034
STD	0.107	0.270	0.799	0.554	0.374	0.750	0.557	0.481	−0.616	−0.564	−0.022
Latent correlations											
MVAL	1.000										
MSC	.586	1.000									
MAFF	.591	.879	1.000								
SVAL	.692	.365	.358	1.000							
SSC	.479	.380	.256	.695	1.000						
SAFF	.463	.284	.329	.713	.822	1.000					
Motivation validity criteria											
MASP	.468	.396	.489	.314	.217	.285	1.000				
SASP	.393	.208	.243	.527	.462	.585	.301	1.000			
MACH	.242	.378	.187	.175	.286	.198	.115	.120	1.000		
SACH	.259	.322	.128	.229	.366	.251	.122	.143	.682	1.000	
EDASP	.168	.216	.136	.172	.203	.127	.132	.104	.163	.152	1.000
Gender differences											
Male	−.060	.093	.038	−.074	.018	−.058	−.037	−.058	−.060	−.080	−.041
SE	.029	.032	.029	.026	.027	.026	.021	.021	.034	.033	.021
United States											
Latent means											
USTD	0.094	−0.263	−0.336	−0.459	−0.335	−0.301	−0.691	−0.687	0.445	0.448	0.228
SE	.032	.032	.038	.034	.033	.039	.033	.034	.023	.025	.018
STD	−0.094	−0.263	−0.336	−0.459	−0.335	−0.301	−0.685	−0.661	0.728	0.643	0.278
Latent correlations											
MVAL	1.000										
MSC	.317	1.000									
MAFF	.449	.713	1.000								
SVAL	.538	.166	.235	1.000							
SSC	.153	.061	.011	.439	1.000						
SAFF	.215	−.010	.160	.543	.741	1.000					
Motivation validity criteria											
MASP	.430	.488	.679	.273	.036	.154	1.000				
SASP	.249	.015	.147	.551	.592	.755	.284	1.000			
MACH	.082	.476	.203	.119	.220	.057	.153	.071	1.000		
SACH	.078	.330	.069	.180	.353	.193	.054	.178	.738	1.000	
EDASP	.288	.275	.207	.267	.225	.175	.205	.187	.313	.312	1.000
Gender differences											
Male	−.041	.092	.002	.007	.102	.081	.003	.063	.025	.075	−.086
SE	.016	.017	.015	.014	.016	.015	.013	.014	.016	.015	.013
Australia											
Latent means											
Mean	−0.423	−0.432	−0.503	−0.894	−0.675	−0.513	−0.905	−0.882	0.373	0.410	−0.576
SE	.039	.041	.045	.047	.040	.048	.036	.040	.038	.038	.040
STD	−0.408	−0.465	−0.531	−0.698	−0.712	−0.484	−0.977	−0.873	0.594	0.613	−0.602
Latent correlations											
MVAL	1.000										
MSC	.392	1.000									
MAFF	.474	.635	1.000								
SVAL	.560	.298	.327	1.000							
SSC	.276	.356	.244	.503	1.000						
SAFF	.324	.215	.390	.561	.784	1.000					
Motivation validity criteria											
MASP	.402	.410	.697	.285	.167	.303	1.000				
SASP	.329	.193	.351	.529	.640	.797	.379	1.000			
MACH	.143	.589	.280	.172	.229	.113	.154	.080	1.000		
SACH	.130	.472	.183	.232	.408	.257	.066	.192	.739	1.000	
EDASP	.252	.387	.287	.328	.317	.256	.209	.227	.444	.459	1.000
Gender differences											
Male	.123	.169	.082	.083	.169	.118	.075	.138	.104	.135	−.035
SE	.023	.027	.028	.023	.025	.023	.021	.021	.044	.043	.031

(*table continues*)

Table 4 (*continued*)

Country/variable	Estimated correlation matrix for the latent variables										
	MVAL	MSC	MAFF	SVAL	SSC	SAFF	MASP	SASP	MACH	SACH	EDASP
England											
Latent means											
USTD	0.503	−0.291	−0.392	−0.580	−0.419	−0.336	−0.886	−0.824	0.528	0.638	−0.519
SE	.037	.038	.043	.040	.038	.043	.033	.034	.042	.041	.048
STD	−0.557	−0.340	−0.437	−0.544	−0.442	−0.339	−1.054	−0.862	0.786	0.909	−0.468
Latent correlations											
MVAL	1.000										
MSC	.346	1.000									
MAFF	.417	.663	1.000								
SVAL	.554	.155	.207	1.000							
SSC	.229	.169	.089	.491	1.000						
SAFF	.257	.077	.253	.557	.778	1.000					
Motivation validity criteria											
MASP	.359	.307	.594	.210	.080	.223	1.000				
SASP	.314	.075	.240	.540	.535	.703	.393	1.000			
MACH	.038	.470	.207	.105	.193	.026	.025	.361	1.000		
SACH	.035	.313	.106	.175	.370	.023	.022	.241	.777	1.000	
EDASP	.159	.244	.153	.217	.197	.108	.103	.187	.442	.447	1.000
Gender differences											
Male	.124	.251	.122	.081	.221	.137	.044	.116	.034	.051	−.079
SE	.026	.022	.019	.023	.021	.022	.021	.021	.031	.031	.030
Scotland											
Latent means											
USTD	−0.210	−0.253	0.498	−0.454	−0.440	−0.288	−0.871	−0.701	0.308	0.256	−0.485
SE	.034	.035	.041	.039	.039	.045	.034	.035	.034	.033	.035
STD	−0.230	−0.305	−0.545	−0.392	−0.429	−0.279	−0.961	0.687	0.493	0.387	−0.521
Latent correlations											
MVAL	1.000										
MSC	.423	1.000									
MAFF	.460	.615	1.000								
SVAL	.421	.217	.227	1.000							
SSC	.251	.294	.143	.581	1.000						
SAFF	.249	.152	.266	.591	.805	1.000					
Motivation validity criteria											
MASP	.356	.270	.606	.216	.089	.213	1.000				
SASP	.282	.135	.254	.567	.609	.734	.360	1.000			
MACH	.111	.456	.147	.195	.363	.205	−.012	.110	1.000		
SACH	.105	.318	.087	.253	.488	.335	−.019	.214	.740	1.000	
EDASP	.150	.256	.138	.229	.334	.230	.098	.202	.479	.507	1.00
Gender differences											
Male	.105	.132	.015	.095	.146	.097	−.022	.077	.009	.043	−.141
SE	.019	.021	.023	.021	.022	.022	.021	.018	.024	.026	.023

Note. These are average results over five multiply imputed data sets. M = math; S = Science; VAL = value; SC = self-concept; AFF = affect; MASP = more math coursework; SASP = more science coursework; MACH = math achievement; SACH = science achievement; EDASP = long-term educational aspiration; USTD = unstandardized; STD = standardized; Male = gender (1 = female, 2 = male).

countries, there is only one small, statistically significant difference in favor of girls and 26 significant differences favoring boys. The largest differences are for the two self-concept factors (median [*mdn*] $r = +.15$; see Table 4). Boys also tend to outperform girls in math achievement (*mdn* $r = +.03$) and science achievement (*mdn* $r = +.06$), but these differences are smaller and many are nonsignificant. In contrast to their performance in specific factors in math and science, girls tend to have higher educational aspirations (*mdn* $r = –.083$). These differences are reasonably consistent across the four Anglo countries, but gender differences in favor of boys tend to be smallest in the United States for both motivation and achievement constructs.

Across the four Arab countries, gender differences are also mostly small for the motivation factors, but these differences tend to favor girls. Differences in favor of girls are largest for the two value factors (*mdn* $r = –.07$; see Table 4) but smallest for the self-concept factors (*mdn* $r = .00$). However, clearly the largest gender differences are the substantially higher scores for girls in both math achievement (*mdn* $r = –.12$) and especially science achievement (*mdn* $r = –.22$). Girls also have significantly higher educational aspirations than boys (*mdn* $r = –.10$). Across the Arab countries, Saudi Arabia is most distinctive in that motivational differences mostly favor boys (seven of eight differences are significant), while achievement and educational aspirations strongly favor girls. In this respect, there is evidence of gender stereotypic differences in motivation constructs in favor of Saudi boys, even though Saudi girls score substantially higher on math and science achievement.

In summary, gender differences in motivation tend to favor boys in Anglo countries, but girls in Arab countries. However, the most

dramatic differences for gender differences are in achievement, where boys are favored to a small extent in Anglo countries but girls are substantially favored in Arab countries. For all countries, girls tend to have higher educational aspirations than do boys.

Correlations among motivation factors (Research Question 7a). We begin with an evaluation of correlations among the six multi-item motivation factors (see Table 4). Although the pattern of correlations is similar within each country, consistent with a priori predictions the correlations are systematically higher for the Arab countries. Within each of the math and science domains, the self-concept and positive affect domains are so highly correlated (.62–.88), particularly in the Arab countries (*mdn r* = .80), as to potentially undermine their discriminant validity.

However, there is clear evidence of domain specificity, in that correlations between the same construct for different domains (e.g., math and science self-concept) never approached 1.0. However, these three correlations are systematically higher for the Arab countries (*mdn r*s = .37–.66; see Table 4) than for Anglo countries (*mdn r*s = .23–.55). The two value factors are most highly correlated in all countries (*mdn r* = .60), while the two self-concept factors are most distinct (*mdn r* = .33).

In summary, the discrimination between math and science is clearly evident for students from all countries, but substantially stronger in the Anglo than in Arab countries. However, the very high correlation between the self-concept and affect found in all countries might call into question the ability of students to distinguish between these two constructs. In the next section, we evaluate whether—despite these high correlations—there is support for the discriminant validity of self-concept and affect constructs in relation to external validity criteria.

Convergent and discriminant validity of motivation factors (Research Questions 7b–7c). Our main test of validity of the TIMSS motivation factors is based on their relations with measures of achievement and wanting to take more coursework in each domain (see Table 4). In relation to the math and science domains, support for convergent and discriminant validity (Research Question 7b) requires that math motivation factors be more highly correlated with math criteria (and less correlated with science criteria) and that science motivation factors be more highly correlated with science criteria (and less correlated with math criteria). The results across all eight countries provide clear support in relation to these tests.

A particular concern with the TIMSS data is the very high correlation between the positive affect and self-concept factors. However, our results provide clear support for the convergent and discriminant validity of these factors. More specifically, across both domains and all countries, achievement was substantially more correlated with self-concept than with positive affect or value. Achievement was substantially correlated with self-concept in the same domain, and the sizes of the correlations are similar across domain and country (*mdn r*s = .48 Arab countries, .45 Anglo countries; see Table 4), but correlations with value (*mdn r* = .20) and positive affect (*mdn r* = .22) were much smaller. Conversely, achievement in one domain is not so substantially correlated with self-concept in the nonmatching domain (*mdn r* = 33).

In contrast to achievement, wanting to take more courses in math and science is more correlated with positive affect than with self-concept or value factors. In each case, these results are domain specific and the sizes of the critical correlations in this pattern of results (shown in bold type in Table 4) are reasonably similar across countries and highly domain specific. Thus, the desire to take more math coursework is substantially correlated with math affect (*mdn r* = .57) but is substantially less correlated with science affect (*mdn r* = .25). Similarly, the desire to take more science coursework is substantially correlated with science affect (*mdn r* = .69) but substantially less correlated with math affect (*mdn r* = .26). Although the desire to take further coursework is less correlated with value and self-concept, even here there is clear evidence of domain specificity. In summary, these results provide good support for the convergent and discriminant validity of the TIMSS motivation scales in relation to both the domain and specific motivation constructs and for the generalizability of the results over the eight countries.

Students' educational aspirations (EDASP in Table 4) are not specific to math or science, so are not relevant in evaluating discriminant validity in relation to domain. Nevertheless, it is important to emphasize that they are positively related to motivation factors in all countries. Interestingly, the correlations tend to be higher for self-concept (*mdn r* = .26) than for value (*mdn r* = .20) and affect (*mdn r* = .16). Thus, whereas plans to pursue more math or science coursework are more highly related to domain-specific positive affect than self-concept, long-term educational aspirations are more highly related to self-concept than positive affect.

Discussion

For nearly two decades, TIMSS has been a primary basis of international comparison of countries in terms of educational achievement in math and science and for benchmarking national performance within the participating countries. Although the primary focus of TIMSS has been on standardized achievement tests in math and science, it has also included motivation constructs in each of the data collections. Here we evaluated support for the construct validity of the TIMSS motivation constructs and compared results for Anglo and Arab countries. Although this is of particular interest in Arab countries, where there has not previously been a rigorous psychometric evaluation of these data, the broader juxtaposition of countries is of general interest to cross-cultural researchers because of the substantial cultural differences (e.g., gender differences in Arab countries, with single-sex school systems and gender-differentiated systems across all ages). Although U.S. schools have increasingly been compared with those from Asian countries (e.g., Liu & Meng, 2010), there has not previously been a rigorous comparison of results from the United States and other Anglo countries with Arab countries based on TIMSS data.

Construct Validity of TIMSS Motivational Measures

We began our study with a critical review of motivational constructs in TIMSS. In marked contrast to the technical sophistication of achievement tests, the theoretical basis of and empirical support for the motivation constructs have been weak.

Cross-national comparisons of manifest means in TIMSS reports, construction of the databases, and most secondary data analyses all require the further assumption that reliability and measurement error are invariant over country, but this assumption clearly was not met. Indeed, all the motivational constructs were substantially more reli-

able in the Anglo countries than in the Arab countries (and other countries in the international sample). These differences undermine the appropriateness of comparisons between countries based on manifest scores not corrected for unreliability—both mean differences between countries and country-level differences in correlations. Such comparisons should be based on latent variable models like those considered here. Particularly problematic are the trichotomized (low, medium, high) motivation scores provided in TIMSS databases that should not be used in secondary data analyses. From this perspective, the use of fully latent variables like those used in the present investigation is critical.

Consistent with a priori predictions, substantial method effects were found to be associated with negatively worded items and the use of parallel-worded items used to measure motivational constructs in TIMSS2007. CFAs that did not take these effects into account failed to fit to the data, and these method effects are not easily incorporated into analyses based on manifest scores, which have been the basis of most secondary analyses with TIMSS. Problems associated with negatively worded items, particularly for children but also for adolescents and adults, have long been known. Also we found, consistent with previous research, substantial method effects associated with the use of parallel-worded items that were controlled with the use of correlated uniquenesses. Potentially, there are important advantages as well as some limitations in the use of parallel items to measure motivational constructs in different academic domains. However, none of these issues has been a focus of research reported in TIMSS technical reports or in the construction of their databases used so widely for secondary analyses.

Once we controlled for a priori method effects, there was good support for the four a priori motivational constructs (self-concept, positive affect, value, and further study) in both domains and all eight countries. However, the comparison of latent means over domain (math vs. science) or over country is based on assumptions of strong measurement invariance (of factor loadings and item intercepts). Tests of these assumptions are a potentially important methodological contribution of the present investigation, as this type of analysis apparently has not been done with TIMSS data. Although there was good support for invariance of factor loadings over domain and country, there was only support for partial invariance of item intercepts over country. Particularly when the number of items per factor is so small (three or four items for each of the multi-item scales), partial invariance provides a weak basis for making comparisons of latent means; this is especially so for the self-concept scale, for which there were also complications in relation to negatively worded items.

It is interesting that there is good support for factor loadings even though there are substantial differences in reliability. This implies that there is not support for strict measurement invariance (i.e., invariance of item uniqueness as well as factor loadings and intercepts). We did not formally pursue this model in the present investigation, as none of our tests of latent variables required strict measurement invariance. Nevertheless, this is an important finding in that it calls into question the appropriateness and validity of analyses based on manifest scores—that have been the basis of nearly all analyses of TIMSS data. These results also demonstrate further the usefulness of latent-variable models that actually require fewer assumptions than traditional analyses of manifest scores.

For multifactor constructs—particularly when measured in multiple domains—discriminant validity is critical for the construct validity and usefulness of the measures. Unless students are able to distinguish reliably between the a priori constructs and there is support for the discriminant as well as convergent validity of their responses, the constructs are unlikely to be useful for applied researchers and policy makers. In particular, the separation of self-concept (competency and expectations of success) and positive affect (interest, enjoyment, and intrinsic motivation) has been an ongoing issue of concern in applied self-concept and motivation research as well as in the theoretical models underpinning this applied research. Thus, for example, previous research, like the TIMSS technical reports, typically has found disattenuated correlations of .7 or higher between self-concept and positive affect. Although we too found high correlations between these two constructs, we also found clear support for their discriminant validity. In particular, consistent with a priori predictions, achievement was substantially correlated with self-concept but not with value or positive affect, while plans to pursue further study were more strongly correlated with positive affect than self-concept or value. Furthermore, these correlations were highly domain specific, providing good support for discriminant validity in relation both to the different motivation constructs and to the two achievement domains. However, long-term educational aspirations were more strongly correlated with self-concept than either value or positive affect. Indeed, educational aspirations were better predicted by self-concept than by objective measures of achievement.

Our focus in the present investigation has been on the validity of motivational constructs within each country, rather than the mean differences between countries. Indeed, the good support for the convergent and discriminant validity of motivational constructs in relation to domain and validity criteria generalized reasonably well across country. Nevertheless, some of the most perplexing results are in relation to latent mean differences, particularly gender differences, and the juxtaposition between motivation and achievement differences in the comparison of the two groups of countries. Hence, we conclude with discussion of these results, speculative interpretations of the findings, and directions for further research.

Country and Gender Differences in Achievement and Motivation Constructs

The juxtaposition of latent mean differences in achievement and motivation factors was perplexing but is consistent with what Shen and Tam (2006) have labeled *paradoxical results*. In particular, students from Anglo countries scored substantially higher than students from Arab countries in terms of academic achievement but had consistently lower scores across all eight motivational constructs. The results, although possibly paradoxical, are highly consistent with a substantial body of self-concept research indicating that in large multinational comparisons, self-concept and achievement are consistently correlated positively at the level of the individual student but consistently correlated negatively at the level of country. The focus of this previous research has been on comparisons of Western countries (particularly the United States) and East Asian countries (particularly Japan and China), where U.S. self-concepts are much higher but U.S. achievement test scores are much lower. However, the findings also generalize to countries where the country-aggregate levels of achievement are below U.S. levels. On this basis, we predicted a priori that the self-concepts would be higher in Arab countries but that achievement test scores were substantially lower in Arab countries. Hence,

our results are consistent with a priori predictions based on this well-established pattern of results. Although there is no conclusive explanation or theoretical explanation of this pattern of results in the research literature, we offer several speculations that may contribute to understanding this phenomenon and to further research.

Well-established frames of reference based on self-concept theory and research provide part of the proposed explanation. There is clear evidence that motivational constructs—particularly self-concept—are substantially influenced by frame of reference effects (Marsh, 2007; Marsh et al., 2008). Thus, students from Arab countries form their self-concepts in relation to other students from their own country, rather than students from the United States and other Anglo countries. In this respect, it is not surprising that achievement is as highly related to motivation constructs in Arab countries as it is in Anglo countries, nor that motivational constructs are not substantially lower in Arab than in Anglo countries. Nevertheless, frame of reference effects do not readily explain why motivational constructs are higher in Arab countries (or why students from Anglo countries have higher self-concepts than students from many Asian countries). This is further complicated in the Arab countries in that educational settings are single-sex, so that the frames of reference are specific to each gender (i.e., boys have little opportunity to compare their performances with girls, or vice versa). Furthermore, Abu-Hilal's (2001) research suggested that the socialization process in the school and family leads Arab students—particularly boys—to be less critical of themselves, so that they have higher self-concepts than might be expected from their objective achievements.

More generally, Shen and Pedulla (2000, p. 237; also see Shen & Tam, 2006) offered a related proposal, suggesting that this pattern may reflect "low academic expectations and standards in low performing countries and high academic expectations and standards in high performing countries." In a related proposal, Marsh et al. (2006) suggested that there are inherent cultural differences in the willingness to express positive things about oneself, particularly in highly evaluative constructs like self-concept. Minkov (2008) suggested that similar results might reflect a cultural value of "monumentalism" related to bipolar constructs of self-enhancement versus self-effacement and self-stability/consistency versus self-flexibility and the need for self-improvement. Using results from a number of cross-national studies of achievement, he showed that at the level of country, monumentalism was positively correlated with positive self-beliefs but negatively correlated with achievement. However, in relation to the present investigation, the United States, Saudi Arabia, and other Arab countries were all high—and did not differ substantially—on monumentalism (based on results from Minkov). In summary, while this juxtaposition of achievement and motivation at the individual and country level is consistent with well-established findings from other research, a full explanation requires further theoretical and empirical research. Particularly fruitful lines of research might include more systematic evaluation of frame of reference models from self-concept research (e.g., Marsh et al., 2008) and cultural value research (e.g., Minkov, 2008) but would also benefit from growing sophistication in the integration of structural equation models of latent constructs based on multiple indicators and multilevel models that simultaneously incorporate latent variables at the level of the individual and country (e.g., Marsh, Lüdtke, et al., 2009).

Gender differences are a particularly interesting aspect of the present investigation and may help to further refine researchers' understanding of the aforementioned paradoxical results, in part because of the gender-segregated nature of schools in Arab countries and society more generally. Based on previous research, we predicted a priori that gender differences in favor of girls would be larger in Arab than Anglo countries, but we left as a research question whether these gender differences in achievement would generalize to academic motivation constructs. In relation to math and science achievement, the results were completely unambiguous. In the Anglo countries, there were small but significant differences in favor of boys for both math and science achievement. In Arab countries, there were consistently moderate or substantial differences in favor of girls for both math and science achievement. Our results suggest that this contrasting pattern of gender differences generalizes over Anglo and Arab countries. Furthermore, other research reviewed earlier (e.g., Abu-Hilal, 2001) has suggested that these differences in achievement favoring girls are not specific to math and science subjects. According to Abu-Hilal, the reason that girls achieve more than boys in Arab schools is because girls spend more time and exert more effort on schoolwork in general—not just in math and science. Indirect support for this suggestion is also evident in the present investigation, in that girls from Arab countries have significantly higher long-term educational aspirations in general, even though their plans to pursue further coursework in math and science are similar to those of boys from Arab countries.

Conclusions

In summary, our research highlights methodological weaknesses in the TIMSS approach to motivation in math and science but provided good support for the construct validity of these motivation measures in relation to achievement, plans to take more coursework in math and science, and long-term educational aspirations. Although there was good support for the a priori factor structure and its invariance over domain and country, the factor structure is complicated by strong negative-item method effects and correlated uniquenesses associated with the use of math and science items with parallel wording. Small, stereotypic gender differences favoring boys are still evident in Anglo countries, but gender differences largely favor girls in Arab countries, which have a strong single-sex education system. The juxtaposition of latent mean differences in achievement and motivation factors was perplexing; students from Anglo countries had substantially higher achievement than students from Arab countries but had substantially lower motivation across all eight math and science factors. Based on these results we encourage researchers to pursue cross-cultural studies of motivation based on TIMSS data but recommend the use of latent variable models like those used here, rather than the manifest motivation scores provided in the TIMSS database. However, we would also encourage TIMSS to develop their student survey more fully, including a richer array of psychosocial constructs, with better testing of their psychometric properties (including the use of negatively worded items), and more appropriate advice to secondary data analysts on the use of these measures.

References

Abu-Hilal, M. M. (2001). Correlates of achievement in the United Arab Emirates: A sociocultural study. In D. M. McInerney and S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (Vol. 1, pp. 205–230). Greenwich, CT: Information Age.

Abu-Hilal, M. M., & Aal-Hussain, A. A. (1997). Dimensionality and hierarchy of the SDQ in a non-Western milieu: A test of self-concept invariance across gender. *Journal of Cross-Cultural Psychology, 28,* 535–553. doi:10.1177/0022022197285002

Abu-Hilal, M. M., & Bahri, T. M. (2000). Self-concept: The generalizability of research on the SDQ, Marsh/Shavelson model and I/E reference model to United Arab Emirates students. *Social Behavior and Personality, 28,* 309–322. doi:10.2224/sbp.2000.28.4.309

Arens, A. K., Yeung, A. S., Craven, R. G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology, 103,* 970–981. doi:10.1037/a0025047

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science, 1,* 164–180. doi:10.1111/j.1745-6916.2006.00011.x

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203. doi:10.1207/s15328007sem1302_2

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22,* 231–240. doi:10.1111/j.1745-3984.1985.tb01061.x

Bøe, M. V., Henriksen, E. K., Lyons, T., & Schreiner, C. (2011). Participation in science and technology: Young people's achievement-related choices in late-modern societies. *Studies in Science Education, 47,* 37–72. doi:10.1080/03057267.2011.549621

Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology, 21,* 149–165. doi:10.1006/ceps.1996.0013

Bruner, J. (1996). A narrative model of self-construction. *Psyke & Logos, 17,* 154–170.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. doi:10.1037/0033-2909.105.3.456

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. doi:10.1080/10705510701301834

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36,* 462–494. doi:10.1177/0049124108314720

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4,* 236–264. doi:10.1177/109442810143004

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. doi:10.1207/S15328007SEM0902_5

Chiu, M. S. (2008). Achievements and self-concepts in a comparison of mathematics and science: Exploring the internal/external frame of reference model across 28 countries. *Educational Research and Evaluation, 14,* 235–254. doi:10.1080/13803610802048858

Chiu, M.-S. (2011, October 3). The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0025734

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.

Deci, E. L., & Ryan, R. M. (1991). A motivational approach to self: Integration in personality. In R. Dienstbier (Ed.), *Nebraska Symposium on Motivation: Vol. 38. Perspectives on motivation* (pp. 237–288). Lincoln: University of Nebraska Press.

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55,* 34–43. doi:10.1037/0003-066X.55.1.34

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346. doi:10.1207/S15328007SEM0903_2

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13,* 440–464. doi:10.1207/s15328007sem1303_6

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326. doi:10.1111/j.2044-8317.1994.tb01039.x

Eccles, J. E., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. I., & Midgley, C. (1983). Expectations, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–145). San Francisco, CA: Freeman.

Eccles, J. S. (1983). Expectancies, values, and academic choice: Origins and changes. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 87–134). San Francisco, CA: Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21,* 215–225. doi:10.1177/0146167295213003

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53,* 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eccles, J. S., Wigfield, A., Harold, R., & Blumenfeld, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development, 64,* 830–847. doi:10.2307/1131221

Feather, N. T. (1982). Expectancy-value approaches: Present status and future directions. In N. T. Feather (Ed.), *Expectations and actions: Expectancy-value models in psychology* (pp. 395–420). Hillsdale, NJ: Erlbaum.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology, *American Psychologist, 56,* 218–226. doi:10.1037/0003-066X.56.3.218

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. doi:10.1146/annurev.psych.58.110405.085530

Hassan, M., & Khailifa, A. (1999). Sex differences in science achievement across 10 academic years among high school students in United Arab Emirates. *Psychological Reports, 84,* 747–757. doi:10.2466/pr0.1999.84.3.747

Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Nonnormal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57,* 327–351. doi:10.1111/j.2044-8317.2004.tb00142.x

Heyman, G. D., & Dweck, C. S. (1992). Achievement goals and intrinsic motivation: Their relation and their role in adaptive motivation. *Motivation and Emotion, 16,* 231–247. doi:10.1007/BF00991653

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage.

Liu, S., & Meng, L. (2010). Re-examining factor structure of the attitudinal items from TIMSS 2003 in cross-cultural study of mathematics self-concept, *Educational Psychology, 30,* 699–712. doi:10.1080/01443410.2010.501102

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. doi:10.1037/1082-989X.7.1.19

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive–developmental phenomena. *Developmental Psychology, 22,* 37–49. doi:10.1037/0012-1649.22.1.37

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early adulthood. *Journal of Educational Psychology, 81,* 417–430. doi:10.1037/0022-0663.81.3.417

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2,* 77–172. doi:10.1007/BF01322177

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810–819. doi:10.1037/0022-3514.70.4.810

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Leicester, England: British Psychological Society.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410. doi:10.1037/0033-2909.103.3.391

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. doi:10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Craven, R. G., & Debus, R. (1999). Separation of competency and affect components of multiple dimensions of academic self-concept: A developmental perspective. *Merrill-Palmer Quarterly Journal, 45,* 567–601.

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the Big-Two factor theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research, 38,* 189–224. doi:10.1207/S15327906MBR3802_3

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390. doi:10.1037/0003-066X.58.5.364

Marsh, H. W., & Hau, K.-T. (2003). Big-fish–little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376. doi:10.1037/0003-066X.58.5.364

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal-external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96,* 56–67. doi:10.1037/0022-0663.96.1.56

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32,* 151–170. doi:10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6,* 311–360. doi:10.1207/s15327574ijt0604_1

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big-Five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471–491. doi:10.1037/a0019227

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44,* 764–802. doi:10.1080/00273170903333665

Marsh, H. W., Martin, A., & Debus, R. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct? In R. Riding, & S. Rayner (Eds.), *Self-perception: International perspectives on individual differences* (pp. 149–170). Westport, CT: Ablex.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16,* 439–476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B., Morin, A. J. S., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. *Journal of Educational Psychology, 103,* 701–732. doi:10.1037/a0024122

Marsh, H. W., Parada, R. H., & Ayotte, V. (2004). A multidimensional perspective of relations between self-concept (Self-Description Questionnaire II) and adolescent mental health (Youth Self-Report). *Psychological Assessment, 16,* 27–41. doi:10.1037/1040-3590.16.1.27

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22,* 366–381. doi:10.1037/a0019225

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350. doi:10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 397–416. doi:10.1111/j.1467-8624.2005.00853.x

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its consequences for young adolescents' course enrollment intentions and performances in mathematics. *Journal of Educational Psychology, 82,* 60–70. doi:10.1037/0022-0663.82.1.60

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. doi:10.1007/BF02294825

Minkov, M. (2008). Self-enhancement and self-stability predict school achievement at the national level, *Cross-Cultural Research: The Journal of Comparative Social Science, 42,* 172–196. doi:10.1177/1069397107312956

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008a). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008b). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of*

Mathematical and Statistical Psychology, 38, 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2008–2011). *Mplus user's guide*. Los Angeles, CA: Authors.

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "×" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science, 22,* 1058–1066. doi:10.1177/0956797611415540

Olson, J. F. (Ed.). (2008). *International database and user's guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66,* 543–578. doi:10.2307/1170653

Pajares, F., & Schunk, D. H. (2005). Self-efficacy and self-concept beliefs: Jointly contributing to the quality of human life. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self-research* (Vol. 2, pp. 95–123). Greenwich, CT: Information Age.

Parsons, J. E, Adler, T., & Meece, J. L. (1984). Sex differences in achievement. A test of alternate theories. *Journal of Personality and Social Psychology, 46,* 26–43. doi:10.1037/0022-3514.46.1.26

Ramirez, M. J., & Arora, A. (2004). Reporting TIMSS 2003 questionnaire data. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical manual* (pp. 309–326). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansome & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373–404). New York, NY: Academic Press. doi:10.1016/B978-012619070-0/50035-0

Renninger, K. A. (2009). Interest and identity development in instruction: An inductive model. *Educational Psychologist, 44,* 105–118. doi:10.1080/00461520902832392

Renninger, K. A., Hidi, S., & Krapp, A. (Eds.). (1992). *The role of interest in learning and development*. Hillsdale, NJ: Erlbaum.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*. Advance online publication. doi:10.1037/a0029315

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. doi:10.1037/1082-989X.7.2.147

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish–little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology, 101,* 403–419. doi:10.1037/a0013838

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55,* 5–14. doi:10.1037/0003-066X.55.1.5

Sharabi, H. (1975). *Introduction to the study of Arab society*. Jerusalem, Israel: Salahueddin.

Shen, C. (2002). Revisiting the relationship between students' achievement and their self-perceptions: A cross-national analysis based TIMSS 1999 data. *Assessment in Education: Principles, Policy, & Practice, 9,* 161–184. doi:10.1080/0969594022000001913

Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigor of mathematics and science: A cross-national analysis. *Assessment in Education: Principles, Policy, & Practice, 7,* 237–253. doi:10.1080/713613335

Shen, C., & Tam, H. P. (2006). The paradoxical relationship between students' achievement and their self-perceptions: A cross-national analysis based on three waves of TIMSS data. In H. Wagemaker (Ed.), *The Second IEA International Research Conference: Proceedings of the IRC-2006, Vol. 1: Trends in International Mathematics and Science Study (TIMSS)* (pp. 43–60). Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

Stevenson, H. W., Chen, C., & Lee, S. (1993). Motivation and achievement of gifted children in East Asia and the United States. *Journal for the Education of the Gifted, 16,* 223–250.

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap*. New York, NY: Simon & Schuster.

Stipek, D. J., & Mac Iver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development, 60,* 521–538. doi:10.2307/1130719

Trautwein, U., & Lüdtke, O. (2007). Predicting global and topic-specific certainty beliefs: Domain-specificity and the role of the academic environment. *British Journal of Educational Psychology, 77,* 907–934. doi:10.1348/000709906X169012

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London, England: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31,* 33–51. doi:10.1177/0022022100031001004

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6,* 49–78. doi:10.1007/BF02209024

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12,* 265–310. doi:10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology, 25,* 68–81. doi:10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs and values from childhood through adolescence. In A. Wigfield, & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 92–120). San Diego, CA: Academic Press. doi:10.1016/B978-012750053-9/50006-1

Wigfield, A., Eccles, J. S., & Pintrich, P. R. (1996). Development between the ages of 11 and 25. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 148–185). New York, NY: Macmillan.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89,* 451–469. doi:10.1037/0022-0663.89.3.451

Wigfield, A., Tonks, S., & Klauda, S. L. (2009). Expectancy-value theory. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 55–75). New York, NY: Routledge/Taylor & Francis Group.

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *Journal of Experimental Education, 72,* 331–346. doi:10.3200/JEXE.72.4.331-346

Received December 12, 2011
Revision received August 1, 2012
Accepted August 6, 2012 ■