Journal of Cross-Cultural Psychology
1–28
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0022022113519858
jccp.sagepub.com
SAGE

The Big-Fish-Little-Pond Effect in Mathematics: A Cross-Cultural Comparison of U.S. and Saudi Arabian TIMSS Responses

Herbert W. Marsh[1,2,3], Adel Salah Abduljabbar[2], Philip D. Parker[1], Alexandre J. S. Morin[1], Faisal Abdelfattah[2], and Benjamin Nagengast[4]

Abstract

This substantive-methodological synergy demonstrates evolving multilevel latent-variable models for cross-cultural data. Using Trends in International Mathematics and Science Study (TIMSS) 2007 data for U.S. and Saudi Arabian eighth grade students, we evaluate the psychometric properties (measurement invariance, method effects, and gender differences) of math self-concept, positive affect, coursework aspirations, and achievement. Extending the studies of the "paradoxical cross-cultural self-concept effect" largely based on U.S.-Asian comparisons, country-level differences strongly favored the United States for achievement test scores, but favored Saudi Arabia for self-concept and aspirations. Latent mean gender differences, of particular interest because of Saudi Arabia's single-sex school system, interacted with country for all constructs. The largest interaction was for achievement test scores; there were no significant gender differences for U.S. students (in coed schools), but in single-sex Saudi schools, Saudi girls performed substantially better than Saudi boys. Consistently with previous (mostly Western) research, but not previously evaluated with TIMSS, in each of the four (2 gender × 2 country) groups all three outcomes (self-concept, affect, and aspiration) were positively influenced by individual student achievement but negatively influenced by class-average achievement (the *Big-Fish-Little-Pond Effect*: BFLPE). BFLPEs were similar in size for boys and girls in coeducational (United States) and in single-sex (Saudi) classrooms.

Keywords

self-concept, frame of reference effects, social comparison processes, Trends in International Mathematics and Science Study, developmental: social, methodology, measurement/statistics

[1]Australian Catholic University, Strathfield, Sydney, New South Wales, Australia
[2]King Saud University, Riyadh, Saudi Arabia
[3]University of Oxford, UK
[4]University of Tübingen, Germany

Corresponding Author:
Herbert W. Marsh, Institute of Positive Psychology and Education, Australian Catholic University, Locked Bag 2002, Strathfield, NSW 2135, Australia.
Email: herb.marsh@education.ox.ac.uk

In this substantive-methodological synergy (Marsh & Hau, 2007), we apply evolving multilevel latent-variable models to address substantively important cross-cultural issues with respect to math self-concept and related constructs of Saudi Arabian students and how they compare with U.S. students, based on the Trends in International Mathematics and Science Study (TIMSS 2007) data.

In quantitative cross-cultural research, many studies apply: (a) confirmatory factor analysis (CFA) and structural equation models (SEMs) using multiple indicators that have traditionally been based on single-level models that ignore the multilevel structure inherent in educational and cross-cultural data, and (b) multilevel models that have traditionally been based on manifest (single) indicators, which ignore measurement and sampling error inherent in psychological data. However, progress has been slow in integrating these two dominant analytical approaches into a single framework in a way that they can be easily implemented in applied research—the focus of the present investigation. Early developments (e.g., Goldstein & McDonald, 1988; McDonald, 1993, 1994; also see Goldstein, 2003) laid the foundation for important advances, but they were not easily implemented with the existing software (e.g., McDonald, 1994; Muthén 1989, 1994). Hence, the major methodological contribution of our study is a demonstration of a new and evolving doubly latent model of contextual effects that has a broad applicability in cross-cultural psychology and psychological research more generally.

The major substantive focus of our study is on the cross-cultural generalizability of the growing body of research in support of the Big-Fish-Little-Pond Effect (BFLPE). As illustrated in Figure 1, the key predictions of the BFLPE are as follows: (a) individual achievement is positively related to self-concept (the brighter I am, the higher my self-concept); (b) but school- or class-average achievement is negatively related to self-concept (the brighter my classmates, the lower my self-concept). Although, as reviewed below, there is considerable support for the BFLPE, the present investigation is apparently the first cross-cultural study of the BFLPE using TIMSS data, the first to focus on comparisons between U.S. and Saudi students, the first to evaluate frame-of-reference effects separately for boys and girls in single-sex Saudi classrooms in comparison with coeducational U.S. classrooms, and the first cross-cultural study to be based on the more proximal frame of reference associated with individual classrooms, rather than the school as a whole.

The juxtaposition of the U.S. and Saudi data is of broad interest to cross-cultural researchers as well as to Saudi researchers. Self-concept and achievement in the United States have been compared particularly in relation to Japan, China, and other East Asian countries, what has been referred to as paradoxical findings (e.g., Shen & Tam, 2008; also see Minkov, 2008; Stevenson, Chen, & Lee, 1993; Stevenson & Stigler, 1992) that we refer to as the *paradoxical cross-cultural self-concept effect*. Specifically, even though academic self-concept is positively related to academic achievement at the level of individual students within each country, at the country level, they are negatively related: U.S. students have substantially higher self-concepts than East Asian students even though their academic achievements are substantially lower. Here, we extend this research to Saudi Arabia students who have much lower levels of achievement according to TIMSS documentation, with particular focus on country-level differences in self-concept. This juxtaposition is also of interest because of the extreme gender-segregated (single-sex) school system in Saudi Arabia and associated gender differences in mathematics achievement, self-concept, affect, and coursework aspirations.

Self-Concept, Achievement and the BFLPE

Self-concept is recognized as a major, and universal, core component of well-being and as a central element in human existence (Bandura, 2006; Marsh & Craven, 2006). Self-concept enhancement is seen as a central goal of education and an important vehicle for addressing the social inequities experienced by disadvantaged groups (see Marsh & Craven, 2006). Recognizing the role of positive self-beliefs across all countries, the Organisation for Economic Co-operation

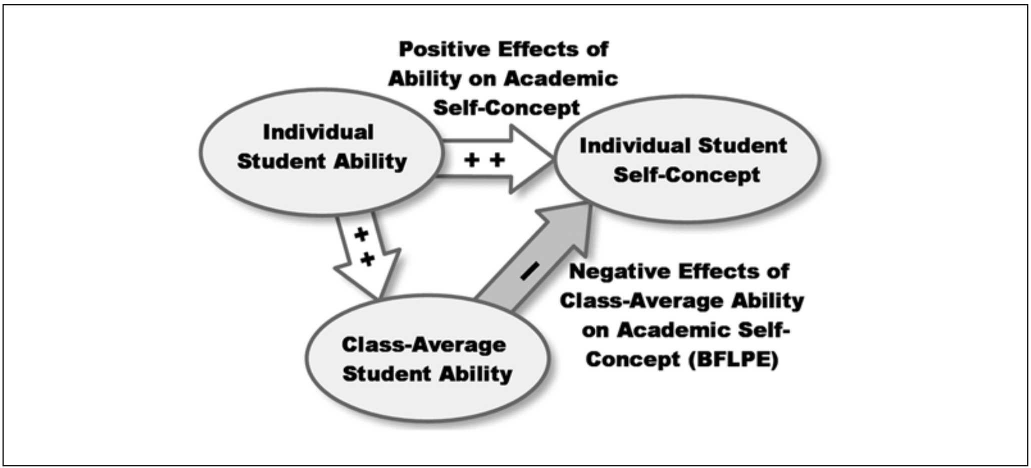
Positive Effects of Ability on Academic Self-Concept
Individual Student Ability
+ +
Individual Student Self-Concept
+
+
–
Negative Effects of Class-Average Ability on Academic Self-Concept (BFLPE)
Class-Average Student Ability

Figure 1. Conceptual path model predictions based on the BFLPE (adapted with permission from Marsh, 2007).
BFLPE = Big-Fish-Little-Pond Effect.

and Development (OECD, 2003) noted that self-concept is "closely tied to students' economic success and long-term health and wellbeing" (p. 9), plays a critical part in students' interest in and satisfaction at school, underpins their academic achievement, and constitutes a very influential platform for pathways beyond school (Ackerman, 2003; Marsh, 2007; Marsh, Hau, Artelt, Baumert, & Peschar, 2006).

In the educational domain, a positive academic self-concept, defined as a positive self-perception of one's academic abilities and competencies (Marsh, Byrne, & Shavelson, 1988) is linked to coursework selection, engagement, intrinsic motivation, subsequent achievement, educational aspirations, and eventual university attendance (e.g., Guay, Marsh, & Boivin, 2003; Marsh, 1991; Marsh & Craven, 2006). For example, Marsh and Yeung (1997a) found that although school grades and self-concept were both substantially correlated with course choice, school grades did not add to the prediction of course choices at all, beyond the substantial effects of subject-specific academic self-concepts. In a longitudinal study over 10 years, Guay et al. (2003) found support for long-lasting beneficial effects of academic self-concept on educational attainment. Academic self-concept and achievement mutually benefit each other (e.g., Marsh & Craven, 2006; Marsh & Yeung, 1997b; Valentine & DuBois, 2005; Valentine, DuBois, & Cooper, 2004); prior achievement has a positive effect on subsequent academic self-concept, and prior academic self-concept has a positive effect on subsequent achievement, even after controlling for prior achievement. Hence, it is not surprising that academic self-concept and related self-belief constructs are a central component in most theories of motivation and social cognition (e.g., Bandura, 2006; Deci & Ryan, 1985; Marsh, 2007; Pajares & Schunk, 2005; Zimmerman, 2008) and fundamentally influence how individuals view the world, the choices that they make, and their subsequent accomplishments. The need to think and feel positively about oneself and the benefits of these positive cognitions on choice, planning, and subsequent accomplishments transcend traditional disciplinary and cultural barriers, and are key ingredients in many psychological theories and are central to goals in many social policy areas.

Self-Concept and Positive Affect

In the TIMSS 2007 database used in the present investigation, students responded to separate scales assessing math self-concept and positive affect (see the appendix), in line with a substantial body of research providing a clear rationale for the separation of self-concept from affective components (e.g., Deci & Ryan, 1985; Eccles, 1983; Eccles & Wigfield, 2002; Feather, 1982; Renninger, 2000, 2009; Renninger, Hidi, & Krapp, 1992; Stipek & Mac Iver, 1989). For example, based on Programme for International Student Assessment (PISA) data from 25 countries, Marsh et al. (2006) reported that even though math self-concept and math interest were very highly correlated (r = .86), math achievement was more highly related to self-concept (r = .33) than to interest (r = .21). However, for the Self Description Questionnaires (SDQ) instruments, which are the basis of much of this research, each academic scale has a combination of competency items (here referred to as academic self-concept) and affect items that are used to assess a single self-concept scale. Based on two large cohorts of students aged 7 to 13, Marsh, Craven, and Debus (1999; also see Arens, Yeung, Craven, & Hasselhorn, 2011) found that the relation between self-concept and affect within the same domain was consistently very high (r = .75), even though self-concept in different domains became more distinct with age (e.g., math and verbal self-concept are substantially correlated for young children but almost uncorrelated by age 12). They lamented that the correlation between self-concept and affect was too high for the constructs to be easily distinguished, but too low for them to be combined into a single factor.

Expectancy-value research (Eccles, 1983; Eccles & Wigfield, 2002) is particularly relevant, showing that correlations between self-concept and affect were evident even for very young children, but increased with age across childhood. Although self-concept and expectations of success, consistent with expectancy-value theory, are typically better predictors of subsequent achievement, Nagengast et al. (2011) reported that affect was better than self-concept in predicting choice behavior (extracurricular activities and career aspirations). In the present investigation, we tested the convergent and discriminant validity of responses to math self-concept and positive affect for Saudi and U.S. boys and girls, hypothesizing that achievement is more strongly correlated with self-concept than positive affect but that coursework aspirations are more correlated with positive affect than self-concept.

Cross-Cultural Generalizability of the BFLPE: Negative Effects of School/Class-Average Achievement on Academic Self-Concept

To understand fully how people perceive themselves, frames of reference must be considered. Depending on the frames of reference or the comparisons individuals use to evaluate themselves, they can reach different conclusions about their accomplishments and so have differing self-concepts as demonstrated in the BFLPE (i.e., academic self-concept is positively predicted by individual achievement, but negatively predicted by class- or school-average achievement; see Figure 1). Thus, the brighter the student, the higher their academic self-concept; but the brighter the student's classmates, the lower their academic self-concept.

The theoretical underpinnings of the BFLPE (see overview by Marsh et al., 2008) lie in theory and research on psychophysical judgment (e.g., Helson, 1964; Marsh, 1974; Parducci, 1995; Wedell & Parducci, 2000), social judgment (e.g., Morse & Gergen, 1970; Upshaw, 1969), sociology (Alwin & Otto, 1977; Hyman, 1942), relative deprivation (Davis, 1966; Stouffer, Suchman, DeVinney, Star, & Williams, 1949), and social comparison (Festinger, 1954). Consistent with the BFLPE predictions, a growing body of research shows that academically selective school systems, ability-grouping, and streaming have detrimental consequences for the academic self-concept of high-achieving pupils. Students in classes or schools with high average achievement have lower academic self-concepts than their equally able peers in schools or classrooms with average or low achievement (e.g., Craven, Marsh, & Print, 2000; Marsh, 1991; Marsh et al., 2008).

BFLPE theory (e.g., Marsh, 1984, 1991; Marsh et al., 2008) posits that the effect is based upon a social comparison process (for a detailed account of the theoretical background of the

BFLPE, see Marsh et al., 2008). Students form their academic self-concepts by comparing their own achievement to the achievement of their classmates, using the class average as a frame of reference. This process tends to have particularly detrimental consequences for high-achieving students in selective academic environments where they mix with other high-achieving students, as it leads to a more negative relativistic perception of their own achievement and more negative academic self-concepts. In less-selective environments, the abilities of peers will have a higher variability and be lower on average, leading to more positive academic self-concepts. On the contrary, students with lower achievement levels can benefit from selective schooling because when they are grouped with other lower achieving students because their frame of reference is based on other low-achieving classmates. Hence, their academic self-concept will be more positive than if they had been placed in a classroom with high-achieving students.

Evidence for this negative contextual effect of school-average achievement on academic self-concept (controlling for the positive effect of individual achievement)—the BFLPE—is strong and has been accumulating for more than two decades (for a comprehensive review, see Marsh et al., 2008). The pervasiveness of the BFLPE has been demonstrated across types of students, academic subjects, and cultures. However, most BFLPE research has been undertaken in Western countries such as Australia (Craven et al., 2000; Marsh, 2004; Marsh, Chessor, Craven, & Roche, 1995; Marsh & Parker, 1984), the United States (e.g., Marsh, 1987, 1991; Mulkey, Catsambis, Steelman, & Crain, 2005), Germany (Marsh, Köller, & Baumert, 2001, Trautwein, Lüdtke, Marsh, & Nagy, 2009), Israel (Zeidner & Schleyer, 1999), France (Huguet et al., 2009; Seaton et al., 2008), the Netherlands (Seaton et al., 2008), and the United Kingdom (Ireson & Hallam, 2009; Ireson, Hallam, & Plewis, 2001; Nagengast & Marsh, 2011; Tymms, 2001), but also in Asian countries (e.g., Liem, Marsh, Martin, McInerney, & Yeung, 2013; Marsh, Kong, & Hau, 2000).

Cross-cultural BFLPE studies. There have been three large cross-cultural studies of the BFLPE (Marsh & Hau, 2003; Nagengast & Marsh, 2011; Seaton, Marsh, & Craven, 2009). Using data from successive waves of PISA data, these studies demonstrated that the BFLPE generalized across different countries and cultures. Summarizing these BFLPE-PISA studies, Nagengast and Marsh (2012) noted that the critical negative effect of class-average achievement on individual self-concept was present in 122 of the 123 samples considered and significant in 114 of them. However, particularly for the earliest of these PISA studies, the countries included were predominantly OECD and Western developed countries and this restricted the generalizability of the findings. Although each of the successive PISA studies included a larger and more diverse sample of countries, Saudi Arabia—the focus of the present investigation—was not included in any PISA study, and Arab countries were highly under-represented in all three studies.

Even though PISA is widely praised as perhaps the best international data available for making international comparisons (e.g., Marsh et al., 2006), it also has strong critics. For example, Hopmann, Brinek, and Retzl (2007; also see Ertl, 2006) summarized a range of important limitations of the PISA data. For instance, they note that the PISA model does not fully reflect the curriculum that is actually taught in many countries, and criticize the fact that PISA measures of mathematics invoke literacy-related abilities. They also note many technical problems related to the sampling design, scaling, translation, and gender, as well as the development of league tables for ranking countries based on PISA. Although we might argue that some, or all, of these problems are well compensated by the incredible richness and scope of PISA data, these multiple limitations clearly indicate the need to cross-validate results based on PISA with data from different sources. Given that the TIMSS data are the major competitor of PISA in terms of international comparisons in mathematics and science, this database is clearly a strong candidate for such replication studies. Although TIMSS and PISA share many similarities, there are also major differences in the achievement tests used in these databases (e.g., American Institutes for

Research, 2005; Hutchison & Schagen, 2007; National Center for Education Statistics, 2008; Neidorf, Binkley, Gattis, & Nohara, 2006; Wu, 2009). Of particular relevance to the preceding criticism of PISA, TIMSS apparently has the relative advantage of assessing achievement in a manner that is more closely related to the academic curriculum. According to Wu (2009), these differences in item content may even partially explain why Western countries tended to perform better on PISA than TIMSS, while the opposite pattern is observed in Eastern European and Asian countries. Importantly, for the purposes of the present investigation, TIMSS does include Saudi Arabia: this allowed comparison of the Saudi results with those from the United States, which has been a primary source of support for the BFLPE and for self-concept research more generally.

Another important difference that is of direct relevance to BFLPE studies is that PISA used schools as the sampling unit, whereas TIMSS uses classrooms. More precisely, in each selected school, PISA tests a random sample of 15-year-olds, so that each school is represented by participants from at least two or more year cohorts. This complicates the interpretation of frame-of-reference effects inherent in the BFLPE since school averages might not correspond to the achievement levels of students in any year cohorts. In contrast, TIMSS targets all students from selected classrooms of eighth grade students (or the year group where a majority of the students are 13 years of age). Selecting the school or the classroom as the frame of reference has been noted as an important consideration for BFLPE research. According to the local dominance effect (Zell and Alicke, 2009; see also Alicke, Zell, & Bloom, 2010), the distinction between the whole school and the individual classroom is a critical issue. They experimentally manipulated "local" and "general" frames of reference in relation to feedback given to participants about how their performances compared with others. Their results show that participants in each condition used the most local comparison information available to them, even when they were told that the local comparison was not representative of the broader population. Consistent with this local dominance prediction in an actual school setting, Liem et al. (2013) showed that track-average achievements for high ability streams within schools were more negatively related to achievement than school-average achievement in the Singapore school system. Based on these results, the classroom sampling unit used in TIMSS seems to represent a far more proximally relevant frame of reference than the school-based sampling used in PISA. In this respect, it is important to note that the present investigation is apparently the first cross-cultural BFLPE study to be based on the classroom as the unit of analysis rather than the school.

Doubly Latent Contextual Effect Models: Substantive-Methodological Synergy

Methodological developments and substantive progress go hand in hand—this is the essence of substantive-methodological synergies (Marsh & Hau, 2007). Coupled with this growing body of BFLPE research are rapidly evolving statistical models of the contextual and climate effects that combine the strengths of latent-variable SEMs and multilevel modeling into an integrated statistical framework (e.g., Marsh et al., 2012) with broad applicability to cross-cultural research. Research on the BFLPE has gradually moved from employing single-level models (e.g., Marsh, 1984, 1987, 1991; Marsh & Parker, 1984) to more appropriate multilevel modeling techniques (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005; Marsh & Hau, 2003; Marsh et al., 2000; Marsh & Rowe, 1996; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Seaton et al., 2008; Seaton, Marsh, & Craven, 2010). Most BFLPE research, however, has either used single-level latent-variable models that ignore the multilevel structure inherent in BFLPE studies, or multilevel models based on manifest indicators that ignore measurement error. It is only recently that applied researchers have been able to combine these two dominant

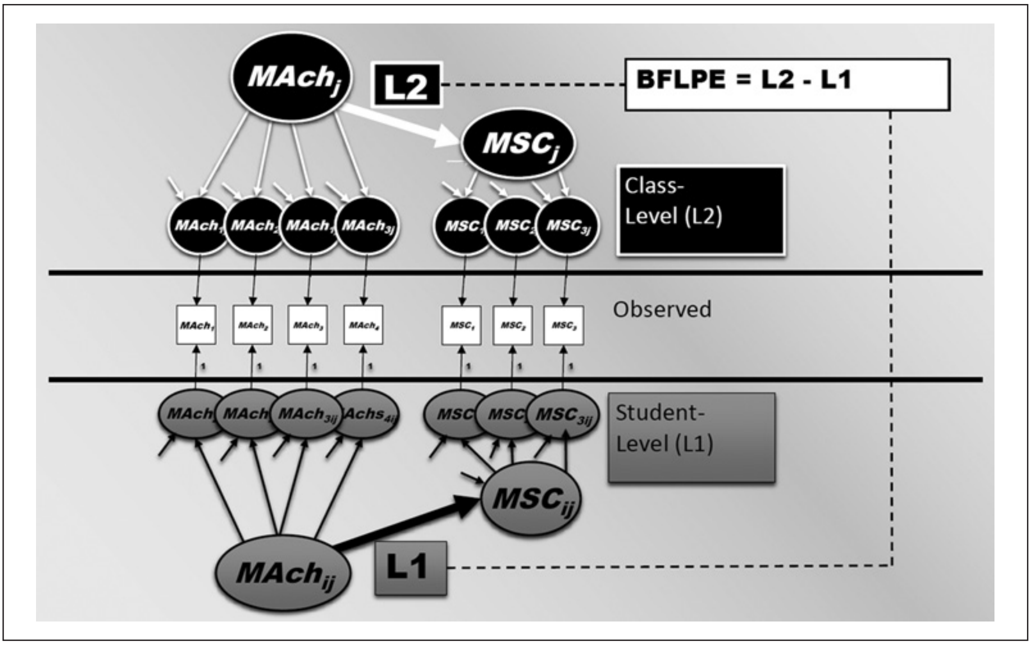

Figure 2. Path diagram of the doubly latent multilevel model.
Note. Subscripts *ij* at the student (L1) level represent the *i*th student in the *j*th class. At the class (L2) level, the subscript *j* refers to the *j*th class. The model is based on implicit-group-mean centering of the L1 variables. Separate indicators of each latent construct are numbered 1 to 4 (math achievement) of 1 to 3 (math self-concept). Paths at L1 represent L1 effects, paths at L2 represent L2 effects. Contextual effects (i.e., the BFLPE in this study) are defined as the difference between L1 and L2 effects. BFLPE = Big-Fish-Little-Pond Effect; MAch = math achievement; L1 = student level; MSC = math self-concept; L2 = class level.

statistical approaches—multilevel modeling and CFA/SEM—into an integrated statistical framework. Consistent with our emphasis on substantive-methodological synergy, these advances have been led by BFLPE research and in turn have made important contributions to it (Marsh et al., 2012; Marsh, Lüdtke, et al., 2009; Nagengast & Marsh, 2012).

Recent developments resulting in stronger statistical models of contextual and climate effects have been led, in part, by BFLPE research (Marsh et al., 2010; Marsh et al., 2012; Nagengast & Marsh, 2012; also see Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Lüdtke et al., 2008). The BFLPE is a contextual effect (see Marsh et al., 2012); the effect of school- or class-average achievement on academic self-concept when the effects of individual student achievement have been controlled (see Figure 2). Class-average achievement, formed by aggregating individual achievement at the student level (L1) to the classroom level (L2), predicts systematic differences in academic self-concept that remain after individual achievement has been controlled. The doubly latent model as implemented in Mplus (Muthén & Muthén, 2008-2013) relies on an implicit group-mean centering of all L1 variables (for further discussion, see Marsh et al., 2012; Marsh, Lüdtke, et al., 2009; Nagengast & Marsh, 2011). For this reason, the effects of L2 variables are not controlled for L1 differences. Hence, estimates of contextual effects are obtained by subtracting the L2 effect from the L1 effect (see Figure 2), which is mathematically equivalent to grand-mean centered results (for further discussion, see Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995; Marsh et al., 2012; Marsh, Lüdtke, et al., 2009).

However, Lüdtke et al. (2008, also see Shin & Raudenbush, 2010) identified a second source of bias in multilevel models—sampling error—that is particularly relevant for the analysis of contextual effects and is not addressed by conventional multilevel modeling techniques (e.g., Hox, 2002; Raudenbush & Bryk, 2002). In the BFLPE, the frame of reference is operationalized as class-average achievement. Although traditional approaches to multilevel modeling treat this class-average achievement as if it were a population value, it is more appropriately viewed as a sample estimate with some degree of sampling error. Drawing upon BFLPE research and the methodological advancements in Lüdtke et al. (2008), Marsh, Lüdtke, et al. (2009; Nagengast & Marsh, 2011) introduced a doubly latent multilevel SEM for contextual effects that controls measurement error at L1 and L2 as well as sampling error in the aggregation of L1 indicators to form L2 indicators (see also Goldstein & McDonald, 1988; McDonald, 1994; Mehta & Neale, 2005; Rabe-Hesketh, Skrondal, & Pickles, 2004). Sampling error can be substantial when agreement among L1 individuals within each L2 class is weak and when the observed sample is both small in size and represents only a small proportion of the population from which the sample was drawn. In this respect, sampling error (in relation to individuals) is similar to measurement error (in relation to items); measurement error can be substantial when the number of items used to infer a latent construct is small and agreement among the different items is weak.

As shown in our representation of the doubly latent model (Figure 2), the observed multiple indicators (represented by boxes) are decomposed into multiple indicators at the individual student (L1) level and at the class (L2) level. In this model, there are CFA models at L1 and L2 controlling measurement error on both levels by using multiple indicators for the considered constructs. These are represented by paths going from each latent construct to the corresponding multiple indicators of that latent construct. Contextual variables (i.e., L2 aggregates of L1 variables) are formed by a latent aggregation procedure that takes into account that the observed values of these variables are not equal to true population values if sampling error is present. Marsh and colleagues showed with a didactic example that the size of the estimated BFLPE could change substantially compared with the standard multilevel regression model and several partial correction models. In this model, estimates of L2 effects are not controlled for L1 effects, so that the contextual effect (i.e., the BFLPE) is defined as the difference between the corresponding L2 and L1 effects (represented as the box labeled as BFLPE = L2-L1 in Figure 2).

A Cross-Cultural Perspective: Generalizability to Saudi Arabia

As described in the TIMSS 2007 Encyclopedia (Mullis, Martin, Olson, Berger, & Stanco, 2008) and elsewhere, Saudi Arabia was founded in 1932 and is geographically the largest Middle Eastern country. It has a population of about 27 million, of which an estimated 16 million are Saudi nationals. From its inception, the governmental system has been monarchic and Islamic. It has the world's largest oil reserves and is the second largest oil exporter, with oil being the basis of most of its exports and the majority of its governmental revenue. Saudi Arabia has a per capita income of about US$24,000, and a substantial budget surplus that, to reduce Saudi dependence on oil, is being invested in economic diversification projects and education.

In 1953, the Ministry of Education was established to make the school system for boys comparable to Western countries; this system was subsequently expanded to include girls. The school system includes 6 years of primary school, 3 years of intermediate school, and 3 years of secondary school. Education is free at all levels. Curricula, teacher training and appointments, and school evaluation are centralized under the Ministry of Education. In addition to providing basic skills and education, schools educate students in beliefs, values, and practices of Islamic culture. The TIMSS Encyclopedia provides an overview of the development of mathematics curricula, indicating that 75% of the curricular content relates to content assessed in the TIMSS tests. A salient feature of the Saudi educational system is that schools are completely segregated in relation to the gender of students and teachers. Consistent with the extreme gender segregation observed throughout Saudi society and the substantial gap between the rights of men and women, students are taught in separate single-sex schools in which male teachers teach boys and female teachers teach girls. The higher education sector has increased substantially since 2000, but also remains almost completely segregated by gender. Based on World Bank 2006 figures, higher education enrollment was over 30% (36.1% for women, 24.7% for men), but these figures are increasing, due to a generous scholarship program that offers tuition and living expenses for Saudi students.

In the Saudi educational system, students take required mathematics classes in first grade (two classes per week), in Grades 2 to 3 (four classes per week), Grades 4 to 6 (five classes per week), then in middle school (four classes per week). Starting in Grades 11 and 12, students can choose to take advanced mathematics coursework as part of the scientific stream if they qualify, but less emphasis is placed on mathematics if students choose a literary or religious stream. At the high school level, course choice for boys and girls is the same (except for physical education for boys and home economics for girls). Although, in terms of university places and post-school employment, not all specializations (e.g., some areas of science and technology, engineering, and agriculture) are open to girls, this imbalance of opportunities has continued to diminish over time.

The United Nations Development Programme (UNDP) provides a set of indices on which human development, poverty, and gender inequality are measured that allow comparison across countries. The most recent ranking provides an interesting comparison between the U.S. and Saudi Arabia (all following results are drawn from UNDP, 2011). On the human development index (HDI), which is a composite of life expectancy at birth, mean and expected years of schooling, and gross national income per capita, the United States is ranked fourth with a HDI value of .910 (higher values are better). Saudi Arabia, while ranked lower than the United States at 56th, is ranked as a high human development country with a HDI value of .770. Furthermore, Saudi Arabia has shown steady increase in this index since 1980. In contrast to the moderate difference between the United States and Saudi Arabia on the HDI, the country rank difference in Gender Inequality Index (GII) is almost twice as large. The United States is ranked 47th with a GII value of .299 (lower values equal more gender equality), whereas Saudi Arabia is ranked 135th with a GII value of .646. Indeed, on the GII Saudi Arabia is the lowest ranked of all medium to high human development countries (as measured by the HDI). The GII is a composite measure with indices of health, education, and political and labor force participation. For the current research, education and labor force participation are most pertinent. In the United States, more women than men over the age of 25 years have a secondary level of education, and the labor force participation rate of women is just over 10% lower than that of men. In Saudi Arabia, 10% fewer women than men over the age of 25 years have a secondary level of education, and there is an almost 50% point different in labor force participation rates between men and women. Nevertheless, in recent years Saudi Arabia has experienced some of the world's largest increases in female education participation ratings, which have risen dramatically since 1975 and are like to continue to increase. Indeed, in recent years, enrollment rates for all levels of education are now at, near, or above parity. The increase in tertiary education for females reflects a worldwide trend, including enrollment and graduation rates that strongly favor females in the United States (OECD, 2011).

Research in Saudi Arabia and in Arab countries more generally has sought to test the cross-cultural generalizability of Western self-concept research findings (e.g., Abu-Hilal, 2001; Abu-Hilal & Aal-Hussain, 1997; Abu-Hilal & Abeld-Mamid, 1989; Abu-Hilal & Bahri, 2000; Marsh et al., 2013). Abu-Hilal and Bahri (2000), in a study of elementary and junior high school students from the United Arab Emirates, noted that self-concept factors were less differentiated (more correlated) across multiple domains than typically found in Western research. They suggested that Arab students (boys in particular) tend to be socialized 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; also see Sharabi, 1975). Thus, Arab students tend to be less aware of their relative strengths and weaknesses than Western students of a similar age, resulting in self-concepts that are more uniformly high and less differentiated. Abu-Hilal (2001) also noted that in adolescence, Arab girls tend to have less freedom than boys, leading them to focus more on schoolwork than boys (also see Abu-Hilal & Abeld-Mamid, 1989; Hassan & Khailifa, 1999). Accordingly, girls had substantially higher verbal and math achievement test scores, were more motivated in school, reported investing more effort in their schoolwork, and had slightly higher math and verbal self-concepts.

TIMSS 2007 documents (Mullis, Martin, & Foy, 2008) show that Saudi students perform much more poorly on the math achievement tests than U.S. students. Although it might be expected that this poorer achievement should lead to lower levels of self-concept, affect and coursework aspirations, the process of forming self-beliefs is complex and is not a simple function of achievement levels, as shown by the *paradoxical cross-cultural self-concept effect* described earlier. According to this effect, there is a positive correlation between self-concept and achievement within countries—a non-paradoxical finding that is well established and supports the construct validity of self-concept (see earlier discussion). The paradoxical aspect of this effect is that at the country level, self-concept and achievement are negatively correlated. The most widely publicized results in support of this effect are comparisons of U.S. and East Asian students (Shen & Tam, 2008; also see Minkov, 2008; Stevenson et al., 1993; Stevenson & Stigler, 1992) showing that U.S. students have substantially higher self-concepts than East Asian students even though their achievements are substantially lower. Extending this research to comparisons between the U.S. and Saudi students, the *paradoxical cross-cultural self-concept effect* predicts that self-concepts should be higher in Saudi Arabia than in the United States even though achievement test scores are higher in the United States than in Saudi Arabia.

Based on the TIMSS 2007 results (Mullis, Martin, & Foy, 2008), we also know, consistent with the evidence that historically observed gender differences favoring boys in math and science achievement are declining, disappearing, or reversing in direction, that international gender differences in math achievement favor girls slightly. However, these data also show that in the United States gender differences in math achievement favor boys slightly, whereas in Saudi Arabia they substantially favor girls. In self-concept research, there are well-established gender stereotypic differences in self-concept, favoring boys in math and science, but girls in verbal areas (Eccles, Adler, & Meece, 1984; Eccles & Wigfield, 1995, 2002; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Marsh, 1989c, 2007). Furthermore, these gender differences in math self-concept persist or even increase after controlling for achievement. Although there is also some speculation that these gender differences are exacerbated by coeducational schooling, empirical evidence does not support this contention (Marsh, 1989a, 1989b; Marsh, Smith, Myers, & Owens, 1988; Nagengast, Marsh, & Hau, 2013). A number of studies show that the BFLPE is not moderated by gender in coeducational schools where both boys and girls are aware of the performances of other students of both genders (see review by Marsh et al., 2008; also see Liem et al., 2013). However, we know of no studies that have evaluated whether the BFLPE varies as a function of gender in single-sex schools where boys do not know how their performances compared with those of girls, and girls do not know how their performances compared with those of boys. Furthermore, the decline in the number of single-sex schools in the United States (and most Western countries) makes these issues difficult to study with U.S. data. From this perspective, the comparison between a completely single-sex school system in Saudi Arabia and the largely coeducational system in the United States is a particularly interesting feature of the present investigation. Thus, we leave as an open research question how gender differences in math achievement in Saudi Arabia and the United States are reflected in gender differences in self-concept, affect, and coursework aspirations in the two countries, and whether the size and direction of the BFLPE vary for boys and girls in the two countries.

The Present Investigation

For the purposes of this substantive-methodological synergy, we focus on relations of math self-concept and positive affect with two validity correlates (math achievement and math aspirations) in four (2 gender × 2 countries) groups based on the TIMSS 2007 data. In preliminary analyses, we tested the a priori factor structure for the self-concept and affect scales, measurement invariance across the four groups that are prerequisite for all subsequent analyses, and the construct validity of responses to self-concept and affect scales in relation to achievement and coursework aspirations.

Based on multigroup tests of measurement invariance over the four groups, we evaluate latent mean differences for country, gender, and their interaction. Of particular interest is the juxtaposition of gender differences in the two countries and the juxtaposition of country-level differences, particularly in achievement and self-concept, but also affect and coursework aspirations, in relation to the paradoxical cross-cultural self-concept effect. From published TIMSS documentation (Mullis, Martin, & Foy, 2008), we already know that Saudi students have much lower math achievement levels than U.S. students. Although it might logically be expected that this would lead to lower math self-concepts for Saudi students, previous research on the paradoxical distance effect suggests that Saudi students might actually have higher math self-concepts. As discussed earlier, Western research suggests that gender differences favoring boys in math achievement have almost disappeared, while differences favoring boys in math self-concepts are still substantial. However, Arab research discussed earlier suggests that Arab girls have substantially higher levels of academic achievement. However, because of the extreme single-sex school system, academic self-concepts of girls are formed largely in relation to accomplishments of other girls, so that higher levels of achievement might not translate into higher academic self-concepts.

Our main focus is to evaluate the generalizability of support for the BFLPE across the four (2 gender × 2 country) groups, with a particular emphasis on gender differences in the size of the BFLPE in single-sex Saudi and coeducational U.S. classrooms. Based on previous research based on PISA data showing that the BFLPE is very robust across countries and gender, we expect that at least the direction of the BFLPE (i.e., the negative effect of school-average achievement) will generalize over gender and country. Concluding, we discuss substantive findings in relation to theory and practice, and offer methodological recommendations for the use of doubly latent models in cross-cultural research.

Method

The TIMSS 2007 data used in the present investigation is the fourth cycle of the TIMSS studies conducted by the International Association for the Evaluation of Educational Achievement (IEA), and includes nationally representative samples of fourth and eighth grade students from 59 participating countries. Although the primary focus of TIMSS has been on the substantive, theoretical, and methodological excellence of achievement tests, TIMSS also administers a student questionnaire of attitudes—the focus of the present investigation. For additional details on the TIMSS 2007, including the development of instruments, translation, sampling, procedures, scaling, and analysis, see Olson, Martin, and Mullis (2008).

Participants and Materials

In TIMSS 2007, the basic sampling strategies follow a two-stage cluster design consisting of sampling of schools from which intact classrooms from the target grade are then sampled (Olson et al., 2008). For the present investigation, participants were eighth grade students from Saudi

Arabia (4,243 students, 52% male, from 203 intact single-sex classrooms) and the United States (7,377 students, 49% male, from 509 intact classrooms, 99.6% coeducational).

Our major focus was on responses to eight items from the TIMSS survey: four math self-concept; three math positive affect; one math coursework aspirations, based on the responses to an agree–disagree Likert-type response scale (see the appendix). The Cronbach's alpha estimates of reliability for math self-concept and affect scales reported in the TIMSS 2007 Technical Manual (Olson et al., 2008) for the two countries varied substantially (see the appendix). Although reliability estimates based on U.S. responses reached an acceptable level of .80 (self-concept, .84; affect, .86), those for Saudi Arabia were much lower (self-concept, .49; affect, .72). For both scales, the median alpha across all participating countries fell between these two values (self-concept, .73; affect, .81). Of particular concern are the unacceptably low reliability estimates for Saudi responses to math self-concept ($\alpha = .49$), suggesting perhaps, problems in the definition of these constructs. The relatively low levels of reliability for Saudi Arabia (and for the international sample more generally) are worrisome, particularly for analyses based on manifest scores. These differences in reliability between Saudi and U.S. responses—but also the differences in reliability for the self-concept and affect constructs—undermine the validity of interpretations based on the manifest scale scores and dictate the use of latent-variable models which provide a natural control for unreliability. Importantly, differences in reliability in latent-variable models are controlled, so that differences in reliability do not undermine interpretation of the results.

Math achievement test scores in the TIMSS database are a mix of constructed responses and multiple choice items from four domains (Algebra, Data and Chance, Number, and Geometry) selected on the basis of item analyses for responses from large-scale pilot studies (Olson et al., 2008). TIMSS 2007 reports the achievement test scores as a set of five plausible values for each student. Plausible values are numbers that are randomly drawn from the distribution of scores that reasonably depict each student's level of achievement. All data analyses were run separately for each of the five plausible values, and the results were aggregated appropriately to obtain unbiased estimates. To take into account the missing data (less than 2% for all survey items and none for achievement test scores), we relied on full information maximum likelihood (Graham, 2009; Schafer & Graham, 2002).

Data Analysis

Analyses were conducted with Mplus 6.1 (Muthén & Muthén, 2008-2013) and consisted of CFA and SEMs based on the Mplus robust maximum likelihood estimator (MLR) and with standard errors and tests of fit that were robust in relation to non-normality and the non-independence of the observations (Muthén & Muthén, 2008-2013). In these analyses, we used the TIMSS's HOUWGT weighting variable, which incorporates six components (sampling of the school, class, and student, and adjustment factors associated with non-participation at the level of school, class, and student). HOUWGT is based on the size of the country-specific samples that is appropriate for the correct computation of standard errors. Previous analyses of TIMSS 2003 (Chiu, 2008, 2011) and 2007 (Marsh et al., 2013) found a method effect associated with negatively worded items, which needs to be controlled by the inclusion of correlated uniqueness. Similarly, consistent across all four (2 gender × 2 country) groups, we found negative-item method effects for the three negatively worded items (see the appendix) that were also controlled by including correlated uniqueness (see Supplemental Materials for further discussion).

Goodness of fit. In applied CFA/SEM research, there is a predominant focus on indices that are sample size independent (e.g., Marsh, Balla & Hau, 1996; Marsh, Balla, & McDonald, 1988; Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004), such as the root mean square error of approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI).

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 less than .05 and .08 reflect a close fit and a minimally acceptable fit to the data, respectively. However, as long as the best fitting model is acceptable, the comparison of the relative fit of nested models imposing more or fewer constraints is more important than the absolute level of fit for any one model. Usually, a decrease in fit for the more parsimonious model remains below .01 for the CFI and .015 for the RMSEA, is taken to reflect reasonable support for the more parsimonious model (e.g., F. Chen, Curran, Bollen, Kirby, & Paxton, 2008; F. F. Chen, 2007; Cheung & Rensvold, 2001, 2002). It should be noted that it is also possible for a more restrictive model to result in a better fit than a less restrictive model based on indices incorporating a penalty for lack of parsimony. However, we note that these are only rough guidelines (Marsh et al., 2004) and that researchers should use an eclectic approach based on the subjective integration of multiple sources of information, including an evaluation of the theoretical conformity of the parameter estimates, common sense, and a comparison of viable alternative models.

Doubly latent contextual effect models. The doubly latent model demonstrated represents new and evolving approaches to the evaluation of contextual and climate effects, stimulated at least in part, by BFLPE research (Marsh et al., 2010, 2012; Nagengast & Marsh, 2012; also see Lüdtke et al., 2011; Lüdtke et al., 2008). As described earlier, the doubly latent model integrates CFAs that control for measurement error at L1 and L2 with multilevel models that appropriately address the hierarchical structure of most educational data (students nested within classes and schools) as well as many cross-cultural studies (persons nested within countries). The BFLPE is a contextual effect (see Marsh et al., 2012), the effect of school- or class-average achievement on academic self-concept when the effects of individual student achievement have been controlled. The doubly latent contextual effect models were estimated in Mplus (Muthén & Muthén, 2008-2013), which relies on an implicit group-mean centering of all L1 variables (for further discussion, see Marsh et al., 2012; Marsh, Lüdtke, et al., 2009; Nagengast & Marsh, 2011). This implies that the partial regression weights associated with L1 variables reflect L1 effects, whereas the partial regression weights associated with L2 variables reflect L2 effects that are not controlled for L1 differences. Estimates of contextual effects, that represent the effect of L2 variables after controlling for L1 differences, can be obtained by subtracting the L1 effect from the L2 effect, which is mathematically equivalent to converting the implicitly group-mean centered results to grand-mean centered results (Enders & Tofighi, 2007; Kreft et al., 1995; Marsh et al., 2012; Marsh, Lüdtke, et al., 2009). Standard errors for confidence intervals and hypothesis tests can be obtained by applying the multivariate delta method (see Raykov & Marcoulides, 2004, for an accessible introduction). Using this approach in combination with the model constraint function in Mplus also allows researchers to decompose estimates into tests of main and interaction effects. Here, we constructed the tests of the main and interaction effects of gender (male vs. female) and country (Saudi Arabia vs. United States) and their interaction in relation to different latent constructs (math self-concept, affect, aspirations, and achievement) and the BFLPE.

The standard error for the contextual effect was obtained using the multivariate delta method (see Raykov & Marcoulides, 2004). In all analyses, effect sizes were calculated according to the recommendations by Marsh, Lüdtke, et al. (2009), with the following formula:

$$\mathrm{ES} = 2 \times \beta \times \frac{\sigma_{\mathrm{pred}}}{\sigma_y}, \qquad (1)$$

where β is the unstandardized regression coefficient, σ_{pred} is the standard deviation of the predictor variable (achievement), and σ_y is the average within-country standard deviation of the outcome variable (self-concept, positive affect, or aspirations), resulting in an effect size (ES) metric that is common across countries. This effect size is comparable to Cohen's *d* (Cohen, 1988).

Measurement invariance. The invariance of the factor structure is an important prerequisite to valid comparisons of results across different countries and groups within countries. Indeed, if the underlying measurement model is fundamentally different in different groups or countries, then it means that the constructs being measured are different across countries and that there is no basis for interpreting observed differences. For instance, in the present study, interpretations of differences in means or relations among constructs across countries are based on the assumptions that the factors are the same across countries (i.e., Saudi and U.S. students). Here, we consider a 2 (country) × 2 (gender) classification of invariance tests. Following Meredith (1993; also see Marsh et al., 2010; Marsh, Muthén, et al., 2009), we evaluated a taxonomy of invariance models beginning with no invariance of any parameters (*configural invariance*), invariance of factor loadings (*weak measurement invariance*), invariance of item intercepts (strong measurement invariance), and invariance of measurement error (*strict measurement invariance).* Marsh et al. (2013) previously conducted related studies of the TIMSS 2007 factor structure and measurement invariance across eight countries and found full support for the invariance of factor loadings and partial support for the invariance of intercepts. Because the models considered here are somewhat different (e.g., the major focus on gender and the BFLPE), and given that these analyses are such an important prerequisite for cross-cultural studies, we replicate and extended these earlier analyses.

However, because these preliminary analyses are not a focus of the present investigation and largely replicate the earlier results, a detailed summary of the results is presented in the Supplemental Materials. Consistent with a priori predictions, there was good support across the four (2 gender × 2 country) groups for negative-item method effects (that were invariant across groups), full invariance of factor loadings, and partial invariance of item intercepts (in which the intercepts of the two negatively worded self-concept items were freed). Consistent with the substantial differences between countries in terms of reliability, already discussed, there was a clear lack of support for item uniqueness (strict invariance) over country, although there was a reasonable support for uniqueness invariance over gender within each country. This non-invariance of measurement error calls into question comparisons between countries based on scale scores, but is not necessary for the comparison of latent means and relations between latent variables, which are the basis of the present investigation (Marsh, Muthén, et al., 2009; Meredith, 1993). On the basis of these preliminary analyses, subsequent analyses are based on the partial strong invariance model (full invariance of factor loadings, partial invariance of item intercepts) described in greater detail in the Supplemental Materials.

In preliminary analyses, we evaluated support for this cross-level invariance, extending the corresponding single-level Model M3d. The fit of the model with factor loading invariance over level (L1-student level and L2 class level) was good (see Table 1 in Supplemental Materials for further discussion; CFI = .954, TLI = .941, RMSEA = .054) and did not differ substantially for less parsimonious models in which these cross-level invariance constraints were not imposed.

Results

Construct Validity of TIMSS Math Self-Concept and Affect Factors

We begin with an evaluation of correlations among the four latent factors to evaluate the construct validity of the self-concept and affect constructs. Consistent with previous research reviewed earlier, the self-concept and affect factors are substantially correlated in all four groups (.702 to .792; Table 1). Although the size of this correlation is sufficiently large to call into question the discriminant validity of responses to these two factors, the correlation is significantly less than 1.0 in all four groups (standard errors (*SE*s) vary from .013 to .032). Although the correlation between math self-concept and affect is high in all four groups, it is significantly smaller in the two U.S. groups (boys, .702; girls, .709) than for Saudi girls (.742) and particularly for Saudi boys (.796).

Table 1. Correlations (and Standard Errors) Among the Constructs for the Four Groups.

	Saudi girls				Saudi boys			
	SC	AF	ASP	ACH	SC	AF	ASP	ACH
SC	1.0			1.0				
AF	0.742 (.032)	1.0			0.796 (.035)	1.0		
ASP	0.300 (.037)	0.660 (.022)	1.0		0.445 (.041)	0.683 (.024)	1.0	
ACH	0.525 (.041)	0.150 (.042)	−0.016 (.035)	1.0	0.589 (.036)	0.142 (.039)	0.049 (.035)	1.0
	US girls				US boys			
SC	1.0				1.0			
AF	0.709 (.013)	1.0			0.702 (.015)	1.0		
ASP	0.450 (.019)	0.645 (.013)	1.0		0.500 (.017)	0.693 (.013)	1.0	
ACH	0.490 (.019)	0.198 (.025)	0.121 (.022)	1.0	0.492 (.018)	0.226 (.025)	0.195 (.024)	1.0

Note. Estimated correlations (with standard errors in parentheses). The results are based on a single-level model with full invariance over factor loadings (for further details, see Model 3d in Supplemental Materials). SC: self-concept; AF: affect; ASP: coursework aspirations; ACH: achievement.

We then pursued tests of the convergent and discriminant validity of these self-concept and affect factors by relating them to two validity correlates: achievement and coursework aspirations. Consistent with a priori predictions, achievement was substantially more highly correlated with self-concept (.490 to .589; see Table 1) than affect (.150 to .226), while coursework aspirations were substantially more correlated with affect (.645 to .693) than self-concept (.300 to .500). In summary, this pattern of correlations among the latent constructs provides strong support for the convergent and discriminant validity of the self-concept and affect constructs, which is consistent over all four (gender-by-country) groups.

Latent Mean Differences

Latent mean country and gender differences. Based on the preliminary tests of measurement invariance (see Supplemental Materials), we are now in a position to evaluate latent mean differences (Table 2). Particularly because of the extreme single-sex nature of the Saudi school system (see earlier discussion), gender differences in the two countries are of special interest. Consistent with the 2 country × 2 gender design, using the Mplus model constraint option we partitioned mean differences into main and interaction effects of country and gender, and simple-main effects of gender differences within each country and of country differences within each gender (in Table 2, under the heading "tests of statistical differences").

Consistent with a priori predictions, gender-by-country interactions are all large and highly significant for self-concept, affect, achievement, and coursework aspirations. However, the largest and most dramatic differences are for achievement, where U.S. students scored substantially higher than Saudi students. However, this main effect of country interacted with gender. Although small gender differences favoring boys in the United States were not statistically significant, Saudi girls scored substantially higher than Saudi boys (see simple main effects of gender in Table 2).

For self-concept, affect, and aspirations, the country-level differences were smaller than those observed for achievement (see main effects of country in Table 2). Nevertheless, effects of country, gender, and their interaction are all statistically significant for all three of these constructs (Table 2). Of particular interest is the juxtaposition between differences in achievement and those in self-concept, affect, and aspirations. In contrast to substantially higher achievement scores in the United States, Saudi students scored higher than U.S. students for all three of these constructs. In Saudi Arabia—again in contrast to achievement scores favoring girls—Saudi boys had

Table 2. Latent Means and Significance Tests of Differences in Country, Gender, and Their Interaction.

	Latent means			
	Saudi Arabia		United States	
	Girls	Boys	Girls	Boys
Scale	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
SC	0.000 (.000)	0.000 (.039)	−0.289 (.033)	−0.159 (.033)
AF	0.000 (.000)	0.154 (.057)	−0.237 (.047)	−0.223 (.048)
ASP	0.000 (.000)	0.192 (.052)	−0.601 (.043)	−0.585 (.044)
ACH	0.000 (.000)	−0.192 (.047)	1.369 (.039)	1.393 (.041)

	Tests of Statistical Differences						
	Main and Interaction Effects			Simple Main Effects: Gender		Simple Main Effects: Country	
	Country	Gender	Interaction	Saudi Arabia	United States	Girls	Boys
Scale	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)	*M* (*SE*)
SC	0.448 (.051)	0.130 (.047)	−0.130 (.047)	0.000 (.039)	0.130 (.026)	0.289 (.033)	0.159 (.036)
AF	0.614 (.069)	0.168 (.063)	0.140 (.063)	0.154 (.057)	0.014 (.026)	0.237 (.047)	0.377 (.045)
ASP	1.377 (.063)	0.208 (.058)	0.176 (.058)	0.192 (.052)	0.016 (.026)	0.601 (.043)	0.776 (.043)
ACH	−2.954 (.066)	−0.168 (.053)	−0.216 (.049)	−0.192 (.047)	0.024 (.020)	−1.369 (.039)	−1.585 (.043)

Note. Estimated latent means (with standard errors in parentheses). The results are based on a single-level model with full invariance over factor loadings and partial invariance of item intercepts across the four (2 gender × 2 country) groups. (For further details, see Model 3d in Supplemental Materials.) Using the model constraints, we then test main and interaction effects of country and gender, followed by simple main effects of gender within each country and of country within each gender. SC = self-concept; AF = affect; ASP = coursework aspirations; ACH = achievement; SE = standard error.

significantly higher scores for affect and aspirations, but did not differ significantly from Saudi girls for self-concept (see simple main effects of gender for Saudi students in Table 2). In the United States, non-significant gender differences favoring boys for achievement are consistent with small non-significant gender differences favoring boys for affect and coursework aspirations. However, even though U.S. boys do not differ from U.S. girls in terms of math achievement, boys score significantly higher than U.S. girls on math self-concept (see simple main effects of gender for U.S. students in Table 2).

In summary, U.S. students had substantially higher achievement scores, but Saudi students had higher self-concept, affect, and aspiration scores. Gender differences in these three constructs were small, but tended to favor boys across both countries. However, the most dramatic gender difference between the two countries was for achievement. Whereas U.S. boys tended to score non-significantly higher than U.S. girls, Saudi girls had substantially higher achievement test scores than did Saudi boys. Nevertheless, these achievement differences in favor of Saudi girls did not seem to translate into enhanced self-concept, affect, or aspirations. In both countries, gender differences in math self-concepts favored boys to a substantially greater extent than would be predicted by gender differences in achievement.

BFLPEs of Class-Average Achievement on Self-Concept, Affect, and Aspirations

The BFLPE predicts that individual student achievement has a positive effect on academic self-concept, but that class-average achievement has a negative effect on self-concept. Substantial support for the BFLPE exists, based primarily on Western and, to a lesser extent, Asian countries.

Table 3. BFLPEs and ESs: Tests of Statistical Significance for the Four (2 Country × 2 Gender) Groups.

	Girls		Boys	
	BFLPE (*SE*)	ES (*SE*)	BFLPE (*SE*)	ES (*SE*)
Saudi Arabia				
SC	−0.216 (.123)	−0.311 (.176)	−0.251 (.085)	−0.396 (.135)
AF	−0.572 (.226)	−0.704 (.278)	−0.588 (.128)	−0.791 (.174)
ASP	−0.433 (.182)	−0.468 (.196)	−0.436 (.121)	−0.515 (.144)
United States				
SC	−0.459 (.060)	−0.476 (.062)	−0.369 (.054)	−0.405 (.059)
AF	−0.379 (.072)	−0.335 (.064)	−0.273 (.063)	−0.255 (.059)
ASP	−0.312 (.070)	−0.242 (.054)	−0.191 (.069)	−0.157 (.056)

	Tests of Statistical Differences		
	Country	Gender	Interaction
	M (*SE*)	*M* (*SE*)	*M* (*SE*)
SC	0.173 (.239)	−0.014 (.229)	−0.156 (.222)
AF	−0.905 (.326)	−0.008 (.354)	0.167 (.345)
ASP	−0.584 (.249)	0.038 (.253)	−0.132 (.255)

Note. ESs were computed in relation to the same total group standard deviations for each group. Tests of statistical significance evaluated the main effects of country and gender, and their interaction, for each of the three dependent variables. BFLPE estimates are the contextual effects of class-average achievement for each group. The results are based on a doubly latent multilevel model of contextual effects with full invariance over factor loadings and partial invariance of item intercepts across the four (2 gender × 2 country) groups, and full invariance of factor loadings over the individual student level (L1) and class level (L2) with latent aggregation of indicators from L1 to L2. (For further details, see discussion of Model 5c in Supplemental Materials.) BFLPE = Big-Fish-Little-Pond Effect; ES = effect size; SC = self-concept; AF = positive affect; ASP = coursework aspirations; SE = standard error.

In the present investigation, we test this BFLPE for apparently the first time in Saudi Arabian boys and girls, compare the size and direction of the effect with those based on U.S. boys and girls, and evaluate the generalizability of the effects on self-concept to those for positive affect and coursework aspirations. Critical issues are the generalizability of the BFLPE over country, gender, and the different constructs. Of particular relevance is the question of whether there are differences between the BFLPE for boys and girls in the coeducation classes in the United States compared with the single-sex classes in the Saudi Arabia.

Results presented earlier (Table 1) showed that the positive correlations of math achievement with self-concept, affect, and aspirations were reasonably consistent across the four (2 gender × 2 countries) groups based on the single-level Model M3d. Here we extend this model into a doubly latent multilevel model to evaluate the additional effects of class-average achievement. In this doubly latent contextual model, the interpretations of contextual effects are facilitated by the invariance of factor loadings over the individual student (L1) and class-average (L2) levels.

In support of the generalizability of the BFLPE, the contextual effects of class-average achievement on self-concept, positive affect, and aspirations were all significantly negative for each of the four (2 gender × 2 countries) groups (Table 3). To evaluate the generalizability of these effects further, we partitioned these effects into tests of the main effects of country and gender, and their interaction. There were no statistically significant differences in the size of the BFLPEs as a function of gender, and this non-effect of gender did not vary according to country. For self-concept, there were no significant differences in the size of the BFLPEs. However, the

BFLPEs were significantly larger for Saudi students than for U.S. students for positive affect ($p < .01$) and coursework aspirations ($p < .05$; see main effects of country in Table 3).

In summary, the results of the present investigation provide very strong support for the generalizability of the BFLPE—the negative contextual effect of class-average achievement. The BFLPE was significantly negative for self-concept (the traditional basis of BFLPE tests), positive affect, and coursework aspirations. In support of the cross-cultural generalizability of the BFLPE, the effects were all significantly negative in tests based on responses by Saudi boys and girls, as well as those based on U.S. boys and girls. Indeed, the negative effects of class-average achievement for affect and aspirations were even more negative for Saudi students than for U.S. students.

Summary and Discussion

The TIMSS studies represent a primary basis for international comparisons and benchmarking countries in terms of educational achievement in math and science. Interestingly, in addition to this primary focus on standardized achievement tests, TIMSS has also focused on self-concept and affective constructs in each data collection. In the present investigation, we evaluated the psychometric properties of TIMSS responses for boys and girls from Saudi Arabia and the United States, compared latent mean differences across the four (2 gender × 2 country) groups, and tested the generalizability of the BFLPE across the four groups and across different constructs (self-concept, positive affect, and coursework aspirations). In particular, the juxtaposition of the U.S. and Saudi data is of broad interest to cross-cultural researchers because of the substantial differences between the two countries (e.g., gender differences in Saudi Arabia, with its single-sex school system and highly 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 U.S. and Saudi results based on the TIMSS data.

Construct Validity of TIMSS Self-Concept and Affect Measures

We started this investigation (see the online Supplemental Materials) with a preliminary evaluation of the psychometric properties of the scales used in TIMSS 2007 to assess the constructs of interest. These analyses demonstrated the full invariance of the factor loadings and the partial invariance of item intercepts across the four groups (2 genders × 2 countries), as well as the need to rely on latent-variable methodologies due to the non-invariance of measurement errors and negative-item method effects. Following these preliminary analyses, we moved to the investigation of the discriminant validity, a critical aspect of the construct validity and usefulness of multifactor constructs. Despite the substantial correlations between self-concept and positive affect, consistent with a priori predictions we found clear support for the discriminant validity of these two constructs; achievement was substantially more correlated with self-concept than positive affect, while plans to pursue further study were more strongly correlated with positive affect than self-concept. Following this preliminary support for measurement invariance and construct validity, we then looked at latent mean differences as a function of country, gender, and their interaction.

Country Differences in Achievement, Self-Concept, Affect, and Aspirations: Paradoxical Cross-Cultural Self-Concept Effects

Our results were also consistent with the *paradoxical cross-cultural self-concept effects* identified by Shen and Tam (2008), as well as with an extensive body of multi-national self-concept research, showing that self-concept and achievement are correlated positively at the student level but negatively correlated at the country level. In particular, Saudi students scored substantially lower than U.S. students in terms of academic achievement but had consistently higher scores for self-concept, positive affect, and coursework aspirations. Previous research in this area has mainly focused on comparisons of U.S. students with those from East Asian countries, showing that U.S. students have higher self-concepts but lower achievement test scores. However, our results show that the findings also generalize to countries with aggregate levels of achievement below the level commonly observed in the U.S.. Thus, the *paradoxical cross-cultural self-concept effect* suggests that Saudi self-concepts would be higher than U.S. self-concepts, even though Saudi achievement scores were substantially lower. Hence, our results are consistent with and extend this well-established pattern of results in a methodologically strong study.

Part of the explanation for this effect lays in well-established frames of reference effects on self-concepts that are the basis of the BFLPE. Saudi students form their self-concepts in relation to other Saudi students and not to U.S. students. These frames of reference effects explain why achievement is as highly related to self-concept, affect, and aspirations in Saudi Arabia as it is in the United States, as well as why these constructs are of a similar level in Saudi Arabia and the United States, even though achievement levels are different. However, these constructs are higher in Saudi Arabia than in the United States, something that cannot readily be explained by these frames of reference effects. The single-sex educational setting of Saudi Arabia further complicates this issue by making the Saudi frames of reference specific to each gender (i.e., boys have little opportunity to compare their performances with girls, or vice versa). Clearly, inherent cultural differences in the willingness to express positive things about oneself, particularly in highly evaluative constructs like self-concept (Marsh et al., 2006), might be at play. According to Shen and Pedulla (2000; also see Shen & Tam, 2008), this pattern may reflect "low academic expectations and standards in low performing countries and high academic expectations and standards in high performing countries" (p. 237). Likewise, Abu-Hilal's (2001) reported that Arab students typically receive uniformly high school marks and less diagnostic feedback about their academic achievement, and that the socialization process leads particularly Arab boys to be less critical of themselves. Hence, they are likely to be less able to objectively assess their specific profiles of strengths and weaknesses, so that their self-concepts are higher than it might be expected. In summary, while our results are consistent with well-established findings from other research, a full explanation requires more research. Particularly fruitful lines of research might be to more fully integrate doubly latent multilevel models of the BFLPE, country-level latent constructs, and cultural value research (e.g., Minkov, 2008).

Gender Differences in Saudi Arabia and the United States

Because of the gender-segregated nature of Saudi schools and society more generally, the gender differences observed in the present study provide important insights into our understanding of the aforementioned paradoxical results. Based on previous research, we predicted that gender differences in achievement would favor girls and be larger in Saudi Arabia than in the United States. However, we left as an open question whether these differences would generalize to self-concept, affect, and aspiration constructs. In relation to math achievement, the results were completely unambiguous. Gender differences interacted strongly with country ($p < .001$) in that Saudi girls had significantly higher levels of math achievement than Saudi boys, while in the United States there were small, non-significant differences in favor of boys. In the TIMSS technical reports (also see Marsh et al., 2013), there are only very small gender differences for achievement, in favor of girls, averaged across all participating countries, but in a number of countries the gender differences in favor of girls are quite substantial. Of particular relevance to our study, the countries with the largest gender differences in favor of girls in math are almost all Middle Eastern Islamic countries (six out of seven). Thus, substantial gender differences in favor of girls seem to generalize across many Islamic countries, a difference that might be explained by the fact that girls tend to spend more time and exert more effort on schoolwork in general (e.g., Abu-Hilal, 2001).

However, gender differences favoring Saudi girls over Saudi boys in achievement were not translated into the corresponding differences in self-concept, affect, and aspirations. We suggest that this pattern of results can be explained by a combination of frame of reference effects, single-sex schools, and perhaps gender stereotypes that remain more firmly entrenched in Saudi Arabia than in the United States. In particular, girls in single-sex classes are only exposed to other girls, so that they have no basis for comparing themselves with boys. Consistent with the frame of reference effects found in many self-concept studies, it is thus reasonable that girls' higher achievement scores are not necessarily translated into higher self-concept, affect, and aspirations. Indeed, consistent with this explanation, Saudi boys and girls had similar math self-concepts. However, this does not explain why Saudi girls scored lower than boys for positive affect and coursework aspirations in mathematics. Although this is beyond the scope of the present investigation, we suggest that these differences are specific to math-related school subjects and continuing gender stereotypes in relation to specific academic subjects that are more deeply entrenched in Saudi Arabia than in the United States. However, we note that in both countries the gender differences in favor of boys for math self-concepts are substantially higher than would be predicted by gender differences in math achievement.

BFLPE Generalizability Across Countries, Genders, and Psychosocial Variables

According to the BFLPE, individual student achievement contributes positively to self-concept, but class-average achievement contributes negatively. A growing body of empirical support demonstrates that the BFLPE is remarkably robust, particularly given the initially counter-intuitive nature of the findings, the controversy that surrounds it, and the important policy implications for academic tracking, streaming, and selective schools (Marsh et al., 2008). Although there is strong cross-cultural support for the BFLPE, based largely on PISA data, and though each successive wave of PISA data has included a larger and more diverse group of countries, the selection of countries is skewed toward Western and OECD countries and, more recently, Asian countries. Of particular relevance to the present investigation, PISA was not conducted in Saudi Arabia. We also note that critics of PISA point to idiosyncrasies in PISA data that might undermine the generalizability of BFLPE cross-cultural research based on PISA, while others point to potentially important differences between PISA and TIMSS. Of particular relevance to the present investigation, frames of reference for PISA data are inferred from the more distal school-average achievement and are further complicated by the fact that the target group of 15-year-olds typically represents two or more year cohorts. In contrast, for TIMSS data these are inferred from the more proximal classroom-average achievement based on all students from intact classrooms (see earlier discussion of the local dominance effect). For these reasons, we undertook the present investigation of TIMSS data to evaluate the BFLPE for boys and girls in Saudi Arabia.

The results of our study are very clear. The effects of class-average achievement were negative in each of the four groups for self-concept, positive affect, and coursework aspirations. For self-concept, the focus of most previous research, there were no significant differences between the size of the BFLPE in Saudi Arabia and the United States. Indeed, for positive affect and coursework aspirations, the BFLPE was actually somewhat larger in Saudi Arabia than in the United States. However, although statistically significant due in part to the large sample sizes, these differences were modest in relation to the corresponding standard errors (i.e., *p*s between .05 and .01 for one effect and slightly less than .01 for the other). We predicted a priori that at least the direction of the BFLPE in relation to math self-concept would generalize over all four (2 gender × 2 country) groups, and there was very strong support for these predictions. We also suggested that the negative effects of class-average achievement would also generalize to positive affect and aspirations, although there is much less research based on these constructs (but see Nagengast & Marsh, 2011); there was clear support for these suggestions. Although not predicted a priori, the finding that the BFLPE for positive affect and aspirations was somewhat larger for Saudi students than for U.S. students warrants further research. Nevertheless, for us, the most important finding was that the negative direction of the BFLPE was consistent across all four (2 country × 2 gender) groups for each of the three constructs (self-concept, positive affect, and aspirations).

A critical limitation of our study and earlier PISA studies is that their cross-sectional design makes causal interpretations more problematic. However, there is a large body of academic self-concept research that addresses these concerns (see Marsh, 2007; Marsh et al., 2008). Quasi-experimental longitudinal studies (e.g., Marsh et al., 2000) show that students' academic self-concept declines when students shift from mixed-ability schools to academically selective schools over time (based on pre-post comparisons), compared with students matched on academic ability who continue to attend mixed-ability schools. Furthermore, there is also support for the BFLPE in true experimental studies in laboratory settings where frames of reference are experimentally manipulated and students randomly assigned to conditions (Zell and Alicke, 2009; see also Alicke et al., 2010; Buckingham & Alicke, 2002). There is support for the BFLPE in studies where achievement is based on tests administered before students began high school (e.g., Marsh et al., 2000). Extended longitudinal studies (Marsh et al., 2000; Marsh et al., 2007) show that the BFLPE grows more negative the longer students attend a selective school and is maintained even 2 and 4 years after graduation from high school. Also, there is good support for the theoretical underpinnings of the BFLPE, as it is largely limited to academic components of self-concept and is nearly unrelated to nonacademic components of self-concept and to self-esteem (Marsh, 1987; Marsh & Parker, 1984). In summary, the BFLPE is a very robust effect.

Appendix

Content and Reliability of TIMSS Constructs Used in This Study

Math Self-concept (SC):
Reliability. SC: SA, .49; US, .84; INT .73.
  I usually do well in math (SCp1)
  Math is harder for me than for many of my classmates (SCn2)
  I am just not good at math (SCn3)
  I learn things quickly in math (SCp4)
Math Positive Affect (Aff).
Reliability. Aff: SA, .72; US, .86; INT, .81)
  I enjoy learning math (AffP1)
  Math is boring (AffN2)
  I Like math (AffP3)
Math Coursework (Crse)
  I would like to do more math in school (single item)
Math Achievement (Ach).
  Composite based on algebra; data & chance; number; geometry
Cluster (class ID; school ID; complex design cluster by class)
Country (country ID; 2 = US; 1 = Saudi)

Note. Reliability estimates are Cronbach's alpha estimates reported in the TIMSS 2007 Technical Manual (Olson, Martin, & Mullis, 2008). Responses to the math self-concept, positive affect, and coursework were all along the same 4-point Likert-type (agree–disagree) response scale. TIMSS = Trends in International Mathematics and Science Study; SC = self-concept; Aff = Affect; SA = Saudi Arabia median international value; US = United States median international value; INT = international.

Acknowledgments

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

Supplemental Materials

Supplemental Materials for the present investigaion are available at http://jccp.sagepub.com/supplemental

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper was made possible in part by a grant from the Australian Research Council (DP130102713, DE140100080).

References

Abu-Hilal, M. M. (2001). Correlates of achievement in the United Arab Emirates: A sociocultural study. In D. M. McInerney & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (pp. 205-208). Greenwich, CT: Information Age Publishing.

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

Abu-Hilal, M. M., & Abeld-Mamid, S. (1989). A comparative study of scores of boys and girls in the preparatory and secondary general examination in the UAE. *Journal of Social Affairs*, *23*, 119-150.

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 Personality*, *28*, 309-322. doi:10.2224/sbp.2000.28.4.309

Ackerman, P. L. (2003). Cognitive ability and non-ability trait determinants of expertise. *Educational Researcher*, *32*, 15-20.

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science*, *21*, 174-177. doi:10.1177/0956797609357718

Alwin, D. F., & Otto, L. B. (1977). High school context effects on aspirations. *Sociology of Education*, *50*, 259-273.

American Institutes for Research. (2005). *Reassessing US international mathematics performance: New findings from the 2003 TIMSS and PISA*. Washington, DC: Author. Retrieved from http://www.air.org/files/TIMSS_PISA_math_study1.pdf

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.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*, 164-180.

Buckingham, J. T., & Alicke, M. D. (2002). The influence of individual versus aggregate social comparison and the presence of others on self-evaluations. *Journal of Personality and Social Psychology*, *83*(5), 1117-1130. doi:10.1037/0022-3514.83.5.1117

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.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504.

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.

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

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). The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology*, *104*, 87-107. doi:10.1037/a0025734

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Craven, R. G., Marsh, H. W., & Print, M. (2000). Gifted, streamed and mixed-ability programs for gifted students: Impact on self-concept, motivation, and achievement. *Australian Journal of Education*, *44*, 51-75.

Davis, J. (1966). The campus as a frogpond: An application of the theory of relative deprivation to career decisions for college men. *American Journal of Sociology*, *72*, 17-31.

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

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: W.H. Freeman.

Eccles, J. S., 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

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.

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

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121-138. doi:10.1037/1082-989X.12.2.121

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*, 619-634.

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: Lawrence Erlbaum.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117-140.

Goldstein, H. (2003). *Multilevel statistical models*. London: Edward Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*, 455-467. doi:10.1007/BF02294400

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*, 124-136.

Hassan, M., & Khailifa, A. (1999). Sex differences in science achievement across ten academic years among high school student in United Arab Emirates. *Psychological Reports*, *84*, 747-757.

Helson, H. (1964). *Adaptation-level theory*. New York, NY: Harper & Row.

Hopmann, S., Brinek, G., & Retzl, M. (2007). *PISA according to PISA*. Vienna, Austria: Verlag.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Huguet, P., Dumas, F., Marsh, H. W., Wheeler, L., Seaton, M., Nezlek, J., . . .Regner, I. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*, 671-710.

Hutchison, G., & Schagen, I. (2007). Comparisons between PISA and TIMSS—Are we the man with two watches? In Loveless, T. (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 227-262). Washington, DC: The Brookings Institution.

Hyman, H. (1942). The psychology of subjective status. *Psychological Bulletin*, *39*, 473-474.

Ireson, J., & Hallam, S. (2009). Academic self-concepts in adolescence: Relations with achievement and ability grouping in schools. *Learning and Instruction*, *19*, 201-213.

Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' self-concepts. *British Journal of Educational Psychology*, *71*, 315-326.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1-21. doi:10.1207/s15327906mbr3001_1

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. A. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal*, *50*, 326-370.

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.

Lüdtke, O., Köller, O., Marsh, H., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, *30*, 263-285.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error-correction models. *Psychological Methods*, *16*, 444-467. doi:10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229.

Marsh, H. W. (1974). *Judgmental anchoring: Stimulus and response variables* (Unpublished doctoral dissertation). University of California, Los Angeles.

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education*, *28*, 165-181.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*, 280-295.

Marsh, H. W. (1989a). The effects of attending single-sex and coeducational high schools on achievement, attitudes and behaviors and on sex differences. *Journal of Educational Psychology*, *81*, 70-85.

Marsh, H. W. (1989b). The effects of single-sex and coeducational schools: A response to Lee and Bryk. *Journal of Educational Psychology*, *81*, 651-653.

Marsh, H. W. (1989c). Sex differences in the development of verbal and math constructs: The high school and beyond study. *American Educational Research Journal*, *26*, 191-225.

Marsh, H. W. (1991). Failure of high ability schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal*, *28*, 445-480.

Marsh, H. W. (2004). Negative effects of school-average achievement on academic self-concept: A comparison of the big-fish-little pond effect across Australian states and territories. *Australian Journal of Education*, *48*, 5-26.

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

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K., Nagengast, B. (2013). Factor structure, discriminant and convergent validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, *105*, 108-128. doi:10.1037/a0029907

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: Lawrence 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*, *102*, 391-410.

Marsh, H. W., Byrne, B. M., & Shavelson, R. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, *80*, 366-380.

Marsh, H. W., Chessor, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, *32*, 285-319.

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.

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 of Developmental Psychology*, *45*, 567-601.

Marsh, H. W., & Hau, K.-T. (2003). Big fish little pond effect on academic self-concept: A crosscultural (26 country) test of the negative effects of academically selective schools. *American Psychologist*, *58*, 364-376.

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*, 151-171.

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.

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: Lawrence 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., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, *38*, 321-350.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*, 337-349.

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., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*, 106-124. Retrieved from http://dx.doi.org/10.1080/00461520.2012.670488

Marsh, H. W., Lütdke, 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.

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.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*, 213-231.

Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept: An application of multilevel modeling. *Australian Journal of Education*, *40*, 65-87.

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.

Marsh, H. W., Smith, I. D., Myers, M. R., & Owens, L. (1988). The transition from single-sex to coeducational high schools: Effects on multiple dimensions of self concept and on academic achievement. *American Educational Research Journal*, *25*, 237-269.

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, *44*, 631-669.

Marsh, H. W., & Yeung, A. S. (1997a). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology*, *89*, 41-54.

Marsh, H. W., & Yeung, A. S. (1997b). Coursework selection: Relations to academic self-concept and achievement. *American Educational Research Journal*, *34*, 691-720.

McDonald, R. P. (1993). A general-model for 2-level data with responses missing at random. *Psychometrika*, *58*, 575-585.

McDonald, R. P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research*, *22*, 399-413. doi:10.1177/0049124194022003007

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation modeling. *Psychological Methods*, *10*, 259-284. doi:10.1037/1082-989X.10.3.259

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

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

Morse, S., & Gergen, K. J. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology*, *16*, 148-156.

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Crain, R. L. (2005). The long-term effects of ability grouping in mathematics: A national investigation. *Social Psychology of Education*, *8*, 137-177. doi:10.1007/s11218-005-4014-6

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

Mullis, I. V. S., Martin, M. O., Olson, J. F., Berger, D. R., & Stanco, G. M. (Eds.). (2008). *TIMSS 2007 encyclopedia: A guide to mathematics and science education around the world* (Vol. 1 and 2). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Muthén, B. (1989) Latent variable modelling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, B. O. (1994). Multilevel covariance structure-analysis. *Sociological Methods & Research*, *22*, 376-398

Muthén, L. K., & Muthén, B. O. (2008-2013). *Mplus user's guide*. Los Angeles, CA: Author.

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the United Kingdom, the UK countries and the world: The big-fish-little-pond-effect for PISA 2006. *Educational Psychology*, *31*, 629-656. doi:10.1080/01443410.2011.586416

Nagengast, B., & Marsh, H. W. (2012). Big fish in little ponds aspire more: Mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. *Journal of Educational Psychology*, *104*, 1033-1053. doi:10.1037/a0027697

Nagengast, B., Marsh, H. W., & Hau, K.-T. (2013). Effects of single-sex schooling in the final years of high school: A comparison of analysis of covariance and propensity score matching. *Sex Roles*, *69*(7-8), 404-422. doi:10.1007/s11199-013-0261-8

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "x" 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

National Center for Education Statistics. (2008). *Comparing NAEP, TIMSS, and PISA in mathematics and science*. Retrieved from http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the national assessment of educational progress (NAEP); Trends in International Mathematics and Science Study (TIMSS); and Program for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). Washington, DC: National Center for Education Statistics, US Department of Education.

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

Organisation for Economic Cooperation and Development. (2003). *Student engagement at school: A sense of belonging and participation*. Paris, France: Author.

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.

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications*. Mahwah, NJ: Lawrence Erlbaum.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190. doi:10.1007/BF02295939

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Application and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, *11*, 621-637. doi:10.1207/s15328007sem1104_7

Renninger, K. A. (2000). How might the development of individual interest contribute to the conceptualization of intrinsic motivation? In C. Sansome & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance*. New York, NY: Academic Press.

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

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

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.

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond-effect: Generalizability and moderation—Two sides of the same coin. *American Educational Research Journal*, *47*, 390-433.

Seaton, M., Marsh, H. W., Dumas, F., Huguet, P., Monteil, J.-M., Régner, I., & Wheeler, L. (2008). In search of the big fish: Investigating the coexistence of the big-fish-little-pond effect with the positive effects of upward comparisons. *British Journal of Social Psychology*, *47*, 73-103. doi:10.1348/01446 6607X202309

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

Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. *Assessment in Education*, *7*, 237-253.

Shen, C., & Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. *Educational Research and Evaluation*, *14*(1), 87-100. doi:10.1080/13803610801896653

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26-53. doi:10.3102/1076998609345252

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.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American soldier: Adjustments during army life* (Vol. 1). Princeton, NJ: Princeton University Press.

Trautwein, U., Lüdtke, O., Marsh, H. W., & Nagy, G. (2009). Within-school social comparison: How students perceive the standing of their class predicts academic self-concept. *Journal of Educational Psychology*, *101*, 853-866. doi:10.1037/a0016306

Tymms, P. (2001). A test of the big fish in a little pond hypothesis: An investigation into the feelings of seven-year-old pupils in school. *School Effectiveness and School Improvement*, *12*, 161-181.

United Nations Development Programme. (2011). *Human development report 2011: Sustainability and equity: A better future for all*. New York, NY: Author.

Upshaw, H. S. (1969). The personal reference scale: An approach to social judgment (L. Berkowitz, Ed.). *Advances in Experimental Social Psychology* 4, 315-370.

Valentine, J. C., & DuBois, D. L. (2005). Effects of self-beliefs on academic achievement and vice versa: Separating the chicken from the egg. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *New frontiers of self-research* (pp. 53-77). Charlotte, NC: Information Age.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*, 111-133.

Wedell, D. H., & Parducci, A. (2000). Social comparison: Lessons from basic research on judgment. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (The Plenum series in social/clinical psychology) (pp. 223-252). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects*, *39*, 33-46. doi:10.1007/s11125-009-9109-y

Zeidner, M., & Schleyer, J. (1999). The big-fish-little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology*, *24*, 305-329. doi:10.1006/ceps.1998.0985

Zell, E., & Alicke, M. D. (2009). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology*, *97*, 467-482.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*(1), 166-183. doi:10.3102/0002831207312909