Marsh, Herbert W., Abduljabbar, Adel Salah, Parker, Philip D., Morin, Alexandre J. S., Abdelfattah, Faisal, Nagengast, Benjamin, . . . Abu-Hilal, Maher M. (2015). The Internal/External Frame of Reference Model of Self-Concept and Achievement Relations: Age-Cohort and Cross-Cultural Differences. *American Educational Research Journal, 52*(1), 168-202.

Abstract

The internal/external frame of reference (I/E) model and dimensional comparison theory posit paradoxical relations between achievement (ACH) and self-concept (SC) in mathematics (M) and verbal (V) domains; ACH in each domain positively affects SC in the matching domain (e.g., MACH to MSC) but negatively in the nonmatching domain (e.g., MACH to VSC). This substantive-methodological synergy based on latent variable models of Trends in International Mathematics and Science Study (TIMSS) data supports the generalizability of these predictions in relation to: mathematics and science domains, intrinsic motivation as well as self-concept, and age and nationality, based on nationally representative matched samples of fourth- and eighth-grade students from three Middle Eastern Islamic, five Western, and four Asian countries (*N*=117,321 students) with important theoretical, developmental, cross-cultural, and methodological implications.

Keywords

internal/external frame of reference mode, substantive-methodological synergy, positive developmental psychology, cross-cultural psychology

Positive self-beliefs are at the heart of the positive psychology revolution focusing on how healthy, normal, and exceptional individuals can get the most from life (e.g., Bandura, 2006; Diener, 2000; Marsh & Craven, 2006; Seligman & Csikszentmihalyi, 2000). Self-concept is also an important mediating factor that facilitates the attainment of other desirable outcomes, such as choice behavior, planning, persistence, and subsequent accomplishments (see Marsh, 1994; Parker et al., 2012; Parker, Marsh, Ciarrochi, Marshall, & Abduljabbar, 2013; Parker, Marsh, Lüdtke, & Trautwein, 2013). There has been substantial improvement in the quality of self-concept research in the past 30 years, largely because of better measurement instruments, theoretical models, quantitative methodology, and research design.

The cornerstone of this resurgence was the classic review article by Shavelson, Hubner, and Stanton (1976), who posited self-concept as a multidimensional hierarchical construct where different facets of academic self-concept (ASC) are substantially correlated and form a single higher-order ASC factor; this is consistent with the positive relations routinely observed among achievements in different school subjects (Marsh, 2007). However, subsequent research revealed that mathematics self-concept (MSC) and verbal self-concept (VSC) in particular were nearly uncorrelated; this led to the Marsh/Shavelson revision (Marsh & Shavelson, 1985). Marsh and Shavelson (1985) posited two higher-order ASC factors (mathematics/academic and verbal/academic), a continuum of core ASC factors ranging from VSC at one end to MSC at the other end, and an ordering of ASCs in other domains along this continuum. This perspective has resulted in increased attention to dimensional comparison processes in self-concept research and more specifically the development of the internal/external frame of reference (I/E) model. The present investigation takes a perspective that is both cross-cultural and developmental, testing the I/E model with matched primary and secondary school samples from 13 diverse countries.

The I/E Model: The Theoretical and Substantive Focus

As long ago as William James (1890/1963), psychologists have recognized the same objective characteristics and achievements can lead to quite different self-concepts, depending on the frames of reference or standards of comparison against which individuals evaluate themselves, and these self-beliefs have important consequences for future choice, behavior, and performance. The two most frequently posited frames of reference are based on social and temporal comparisons (Albert, 1977; Möller, 2005; Möller, Pohlmann, Köller, & Marsh, 2009; Möller, Retelsdorf, Köller, & Marsh, 2011). Self-perceptions based on how current accomplishments compare with past performances reflect temporal comparisons, whereas those based on how accomplishments compare with the accomplishments of others reflect social comparisons. Particularly in educational settings, a growing body of research based on the I/E frame of reference model (Marsh, 1986; Marsh et al., in press; Möller & Marsh, 2013) demonstrates that self-perceptions may also be the result of internal or dimensional comparisons, in which accomplishments in one school subject can serve as a frame of reference for another school subject (e.g., Möller et al., 2009; Möller & Köller, 2001).

Initially, the I/E model was developed to explain why MSC and VSC are nearly uncorrelated even though achievement in the same areas are strongly correlated (see Marsh, 1986, 2007; Marsh et al., in press; Möller & Marsh, 2013); people think of themselves as primarily a verbal person or a mathematics person, but rarely both, even though persons good at one also tend to be good at the other (Marsh, 1986, 2007). The I/E model posits that ASC in a particular school subject is formed in relation to two frames of reference: an external (social comparison) reference in which students contrast their perceived performances in a particular school subject with the perceived performances of their peers in the same school subject and an internal (dimensional comparison) reference in which students contrast their own performances in one particular school subject against their own performances in different school subjects.

Tests of the classic I/E model typically focus on mathematics and verbal domains, relating mathematics and verbal achievements to MSC and VSC (see Figure 1). According to the external comparison process, good mathematics skills lead to higher MSCs and good verbal skills lead to higher VSCs. However, the internal comparison process predicts that good mathematics skills lead to lower VSCs after controlling for the positive effects of good verbal skills. In empirical tests of the I/E model (Figure 1), the horizontal paths leading from mathematics achievement to MSC and from verbal achievement to VSC are predicted to be substantially positive (“++” in Figure 1). However, the cross-paths leading from mathematics achievement to VSC and from verbal achievement to MSC (the grey lines in Figure 1) are predicted to be negative (“–” in Figure 1). In a review of I/E studies, Möller et al. (2009; also see Marsh, 2007; Möller & Marsh, 2013) note that evidence in favor of this model comes from diverse sources, based on qualitative introspective and quantitative cross-sectional, longitudinal, quasi-experimental, and true experimental designs.

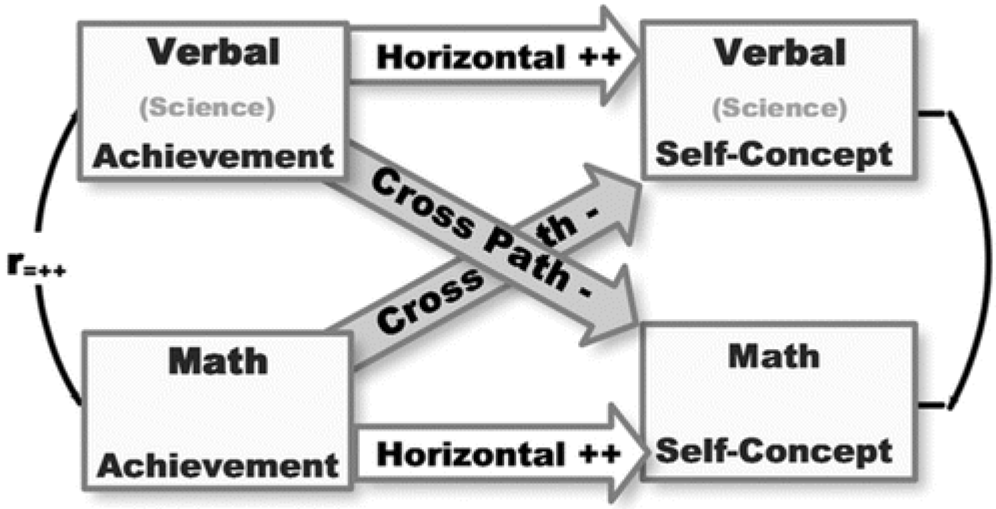

Figure 1. The "classic" internal/external frame of reference (I/E) model relating verbal and mathematics achievement to verbal and mathematics self-concepts. According to predictions from the I/E model, the horizontal paths from achievement to self-concept in the matching domain (content area) are predicted to be substantial and positive (++), whereas the cross-paths from achievement in one domain area to self-concept in a nonmatching domain (shaded in grey) are predicted to be negative (–). In the present investigation we evaluate the generalizability of these predictions to the science domain, relating science and mathematics achievement to science and mathematics self-concepts.

Meta-Analysis Support for the Universality of Predictions Based on the I/E Model

Of particular relevance to evaluating the generalizability of support for I/E predictions is the Möller et al. (2009) meta-analysis, which tested the I/E model in each of 69 independent data sets (*N* = 125,308 students). As predicted by the I/E model, their results revealed positive paths from achievement to corresponding self-concepts (horizontal paths in Figure 1; .61 for mathematics, .49 for verbal) and negative paths from achievement in one subject to self-concept in the other subject (cross-paths in Figure 1; −.21 from mathematics achievement to VSC, −.27 from verbal achievement to MSC). Support for the I/E predictions generalized across age groups, gender, and countries, leading the authors to conclude, "The results of our meta-analyses indicate that the relations described in the classical I/E model are not restricted to a particular achievement or self-concept measure or to specific age groups, gender groups, or countries" (p. 1157). Although there was significant study-to-study variation in the sizes of relations among I/E variables, remarkably, they found no significant differences for country of residence or age. A particular strength of meta-analysis is the ability to evaluate the generalizability of the results, but the strength of these tests depends on the available data. Important limitations include the representativeness of samples and comparability of measures across available studies. In particular, there were important gaps in the available studies with a substantial underrepresentation of young children and countries other than Western and Asian countries. Hence, notable strengths of the Trends in International Mathematics and Science Study (TIMSS) data used here are nationally representative samples of primary and secondary students using carefully constructed, consistent measured for a diverse set of countries. Thus, the overarching purpose of our study is to test more fully the conclusions that support for the I/E model generalizes over age and country, using stronger data and statistical methodology.

Developmental Support for the Generalizability of the I/E Model

For many developmental, educational, and psychological researchers, self-concepts are a "cornerstone of both social and emotional development" (Kagen, Moore, & Bredekamp, 1995, p. 18; also see Davis-Kean & Sandler, 2001; Marsh, Ellis, & Craven, 2002); self-concepts develop early in childhood, and once established, they are enduring (e.g., Eder & Mangelsdorf, 1997). Hattie (1992; Hattie & Marsh, 1996; also see Eccles, Wigfield, Harold, & Blumenfeld, 1993; Harter, 2006, 2012; Marsh, Craven, & Debus, 1998) reviewed theoretical and empirical support for stages of growth in the development of self-concept, arguing against the notion of fixed stages that all persons must pass through. Indeed, many authors (Chapman & Tunmer, 1995; Eccles et al., 1993; Harter, 1999; Marsh, 1989; Marsh et al., 1998, 1999; Skaalvik & Hagtvet, 1990; Wigfield & Eccles, 1992; Wigfield et al., 1997) have offered a developmental perspective on the relation between ASC and academic achievement. For example, Marsh (1989, 1990b; Marsh et al., 1998) proposed that the self-concepts of very young children are very positive and are not highly correlated with external indicators (e.g., skills, accomplishments, achievement, self-concepts inferred by significant others) but that with increasing life experience, children learn their relative strengths and weaknesses so that specific self-concept domains become more differentiated and more highly correlated with external indicators. Marsh et al. (1998) showed that reliability, stability, and factor structure of self-concept scales improve with age (children 5–8 years of age). In addition, consistent with the proposal that children's self-perceptions become more realistic with age, self-ratings of older children were more correlated with inferred self-concept ratings by their teachers.

Based on this developmental theory it seems reasonable to predict that support for the I/E would increase with age, particularly in relation the size of the horizontal paths but also the differentiation between domains that drives the cross-paths. However, there is little direct empirical support or even good tests of the generalizability of support for the I/E model over age. Surprisingly, Möller et al. (2009) found no differences as a function of age given developmental models of how self-concept formation varies as a function of age (e.g., Eccles et al., 1993; Harter, 2012; Marsh, 1989, 2007; Wigfield & Eccles, 2002). Nevertheless, because of the paucity of available studies with young children in this meta-analysis (only 3 of 69 samples reported results for children in Grade 4 or younger), the generalizability of this finding was not strong. Hence, a central purpose of the present investigation is to provide stronger tests of the generalizability of the I/E predictions over primary and secondary school students.

Cross-Cultural Generalizability of Support for the I/E Model

Cross-cultural comparisons provide researchers with a valuable heuristic basis to test the external validity and generalizability of their measures, theories, and models. Matsumoto (2001) argued that: "Cultural differences challenge mainstream theoretical notions about the nature of people and force us to rethink basic theories of personality, perception, cognition, emotion, development, social psychology, and the like in fundamental and profound ways" (p. 9). In cross-cultural research there are two main orientations, one that focuses on tests of a priori hypotheses of cross-cultural differences and one that tests the replicability of existing theories in other cultures and seeks universal, pan-human theories (e.g., Marsh, Hau, Artelt, Baumert, & Peschar, 2006; Parker et al., 2012; Segall, Lonner, & Berry, 1998).

However, there exists a schism between the overarching cultural relativist and universalist perspectives of cross-cultural research (Kagitcibasi & Poortinga, 2000). The broad cultural relativist (idiographic, emic, indigenous, qualitative) perspective emphasizes the uniqueness of the individual case that defies comparison. In contrast, the broad universalist (nomothetic, etic, positivist, quantitative) perspective emphasizes what is common between cultures with an emphasis on theoretical predictions, replicability of results, and empirical testing. In their taxonomy of cross-cultural research, Van de Vijer and Leung (2000)discussed generalizability studies with a strong theoretical framework for generating testable hypotheses and an emphasis on the universality of structures and theoretical propositions. Within this context, they noted the need to use multiple group modeling approaches that allow researchers to make fine-grained comparisons of factor structures and patterns of relations between multiple constructs in different cultural groups. In this framework, there is a focus on similarities as well as a consideration of observed differences. Because of the traditional focus on null hypothesis testing, there is an unfortunate tendency to provide elaborate interpretations for (sometimes very small, idiosyncratic) differences and largely to ignore similarities that may argue for cross-cultural generalization. Van de Vijer and Leung emphasized that the endemic problems of replicability in cross-cultural research will improve with greater emphasis on theory development and testing, coupled with the more appropriate use of new statistical tools. Particularly in education, an ongoing challenge in cross-cultural research is to disentangle the potentially confounding effects of differences in the appropriateness of psychological measures used in different cultural settings, differences in the typically ad hoc samples of participants representing different cultural groups, and true cultural differences.

More broadly conceived, this universalist perspective of cross-cultural comparisons fits into the general “multiple method” approach to construct validity (Marsh, Martin, & Hau, 2006) in which convergence is evaluated across results from different methods—the different countries in this application. To the extent that a priori predictions based on a strong theoretical model generalize reasonably well across responses from diverse set of countries, there is strong support for the construct validity of interpretations based on the theoretical model. Although not denying the relevance of the cultural relativist perspective, the focus of the present investigation is on the universalist perspective and cross-cultural support for the generalizability of I/E predictions across different countries.

Strong cross-cultural studies need to compare the results from at least two—and preferably many—countries based on comparable samples and the same measures; otherwise, apparent cross-cultural differences are confounded with potential differences in the composition of samples and perhaps the appropriateness of materials. Addressing these challenges, there is strong support for the cross-cultural generalizability of the I/E model based on the Organisation for Economic Cooperation and Development (OECD) Program for International Student Assessment (PISA) data; 103,558 15-year-old students from 26 countries (Marsh & Hau, 2004). Across the 26 countries, the two horizontal paths relating mathematics achievement to MSC (.44) and verbal achievement to VSC (.47) were substantial and positive, while the two cross-paths leading from verbal achievement to MSC (−.20) and mathematics achievement to VSC (−.26) were negative. They also noted that the correlation between mathematics and verbal achievements (r = .78) was very large, whereas the corresponding correlation between MSC and VSC (r = .10) was close to zero. Subsequent analyses showed that these results generalized well across all 26 countries.

In their meta-analysis, Möller et al. (2009) incorporated results from the Marsh and Hau (2004) PISA study, representing 26 of their 69 samples. Consistent with conclusions by Marsh and Hau, they found good cross-cultural support for the generalizability of the I/E model in that there were no significant differences for country of residence. However, most of the samples, aside from a few Asian countries, were based on Western countries (7 Asian, 12 Australian, 7 U.S., and 39 European samples were included in the analysis of country as a moderator). Although the meta-analysis included responses by students from "Other Nationalities," the authors argued that the number of studies from these "other" countries was too small to be included in analyses of country as a moderator. However, Schwartz and Bilsky (1990), as well as many others, observed, "Theories that aspire to universality . . . must be tested in numerous, culturally diverse samples" (p. 878). In this respect, one purpose of our study is to greatly expand the scope of tests of the cross-cultural generalizability of the I/E beyond the Marsh and Hau PISA study and Möller et al. meta-analysis that have been the primary basis of cross-cultural support for the universality of support for I/E predictions.

Generalizability of Support for the I/E Model to Middle Eastern Islamic Countries

The generalizability of Western self-concept research findings to Middle Eastern Islamic countries[1] has been the focus of a number of studies by Abu-Hilal and colleagues (Abu-Hilal, 2001; Abu-Hilal & Aal-Hussain, 1997; Abu-Hilal & Abeld-Mamid, 1989; Abu-Hilal & Bahri, 2000; also see Marsh et al., 2013). Abu-Hilal and Bahri (2000) evaluated the generalizability of responses to Marsh's Self Description Questionnaire (SDQ) by elementary and junior high school Arab students. They found support for the a priori factor structure but noted that the ASC factors tended to be less correlated with corresponding areas of achievement and less differentiated (more correlated) than typically is found in Western research. For example, MSC and VSC scales on the SDQ are typically almost uncorrelated (Marsh, 2007). In contrast, for the older junior high school sample, these two ASC scales were moderately correlated (r = .37), but less so than the substantial correlation between corresponding measures of achievement (r = .62). Abu-Hilal and Bahri (2000; also see Sharabi, 1975) noted that Arab students are 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). When they asked Arab students to evaluate their skills and performances in different school subjects, several students commented "Are you sure you want us to judge our performance? I think that teachers can tell you better than we can" (p. 320). Abu-Hilal and Bahri (2000) noted that this pattern of results is similar to that found with younger children in Western research (e.g., Marsh, 1989, 1990a, 1990b), in which self-concepts of young children are also uniformly high and substantially correlated but become more differentiated with age as children obtain more experience relative to their related strengths and weaknesses. In the present investigation we translate these observations into testable hypotheses in relation to I/E predictions and provide more rigorous tests of their validity and of their developmental and cross-cultural generalizability.

Academic Domain

Most studies testing the classic I/E model have investigated the mathematics and verbal (native language) domains. However, subsequent research has tested the generalizability to academic subjects other than the verbal and mathematical domains (e.g., Bong, 1998; Chiu, 2008; Dickhäuser, 2003; Marsh, Kong, & Hau, 2001; Möller, Streblow, Pohlmann, & Köller, 2006; Nagy, Trautwein, Baumert, Köller, & Garrett, 2006; Yeung, Lee, & Wong, 2001). Thus, Möller et al. (2009; also see Marsh et al., in press) called for an extension of the I/E model to other academic domains, asking, for example, whether:

> students see physics and mathematics as sufficiently distinct that better performances in one would lead to poorer self-concepts in the other (a contrast effect like that posited in the I/E model based on the mathematics and verbal domains), or would the two be seen as sufficiently similar so that better performance in one would lead to better self-concepts in the other (an assimilation effect)? (p. 1159).

In the present investigation, we take up this challenge using the mathematics and science constructs measured as part of the TIMSS cross-national study of representative samples of primary (Grade 4) and secondary (Grade 8) students from around the world.

Theoretical Extensions of the I/E Model: Intrinsic Motivation

In the present investigation, we also extend the I/E model to evaluate its generalizability to intrinsic motivation that is driven by an interest, enjoyment, or positive affect associated with the task itself, rather due to external pressures or as a means to external rewards (Deci & Ryan, 1985; Eccles et al., 1983; Eccles & Wigfield, 2002; Renninger, 2009). However, this research typically has focused on intrinsic motivation and its relations to other constructs within a single domain rather than on the juxtaposition of intrinsic motivation in different domains. Expectancy-value theory (EVT) research is particularly relevant. It has shown that correlations between expectancy (typically operationalized as self-concept responses) and interest were evident even for very young children (Eccles et al., 1983; Eccles & Wigfield, 1995; Wigfield & Eccles, 2002; Wigfield, et al., 1997) and increased with age during early school years but that both expectancy and value constructs were highly domain specific (Eccles et al., 1993). Although the domain specificity of these different constructs tends to increase with age, the relations between expectancy and value within the same domain remain high or even increase with age (e.g., Eccles et al., 1993; Eccles & Wigfield, 2002; Wigfield & Eccles, 2002; Wigfield, Eccles, & Pintrich, 1996; also see Marsh et al., 1999). Extending EVT research to incorporate the dimensional comparison perspective of the I/E model, Eccles (2009; also see, Denissen, Zarret, & Eccles, 2007) emphasized "that both external and internal comparison processes are key—people assess their own skills by comparing their performances with those of other people and with their own performances across domains" (p. 82) and demonstrated how this internal dimensional comparison process has a critical influence in academic choice behavior (e.g., choice of university major; see Parker et al., 2012).

Pekrun and colleagues (Goetz, Frenzel, Pekrun, & Hall, 2006; Pekrun, 2006) extended their control value theory of academic emotions to incorporate the I/E model to explain the extreme domain specificity of academic emotions, speculating that "the mechanisms addressed by Marsh's (1986) I/E model could operate for students' emotions as well" (Goetz et al., 2006, p. 7). They found that emotions were substantially more domain specific than achievement in different mathematics and verbal subjects and that enjoyment was the most domain-specific emotion. Goetz, Frenzel, Hall, and Pekrun (2008) subsequently found support for the I/E model in relation to both self-concept and enjoyment but showed that achievement/enjoyment relations were mediated by self-concept. Goetz et al. (2008) also suggested that further research was needed to evaluate relations between mathematics and science constructs.

Following from these studies of expectancy-value theory and control value theory, as well as the Möller et al. (2009) meta-analysis, in the present investigation we extend the I/E model to incorporate intrinsic motivation, evaluate developmental hypotheses about the role of dimensional comparisons in the formation of self-concept and intrinsic motivation, and juxtapose the relations between mathematics and science constructs.

TIMSS 2007: Background to the Present Investigation

Tests of the I/E Model With the TIMSS 2007 Data

In research particularly relevant to the present investigation, Chiu (2012; also see Chiu, 2008) conducted tests of the I/E model for mathematics and science constructs using TIMSS 2007 data for only the eighth-grade cohort. Across all countries there was good support for I/E predictions, leading Chiu to conclude: "Mathematics and science can be distinctly different school subjects perceived by students through the psychological process of internal comparison in constructing their self-concepts in the two domains" (p. 102). Clearly these results support the extension of the I/E model to juxtapose mathematics and science constructs. Nevertheless, support for the I/E model—particularly the negative cross-paths (see Figure 1)—was clearly weaker than previously reported in the Möller et al. (2009) meta-analysis or the Marsh and Hau (2004) cross-national PISA study. More specifically, both cross-paths were significantly negative in less than half the countries considered, and the mean size of the cross-paths was only about −.10 (see Chiu, 2012, Table 4, p. 96) compared to about −.25 in previous I/E studies (e.g., Marsh & Hau, 2004; Möller et al., 2009).

An obvious possible explanation for this difference could be that the I/E model is stronger when based on two domains at opposite ends of the Marsh-Shavelson continuum of academic domains (i.e., mathematics and verbal) than for two domains closer together on this continuum (i.e., mathematics and science). However, some alternative explanations also warrant consideration. In particular, due to the complicated nature of the research questions pursued, Chiu (2012) used single manifest scores based on the trichotomized scale scores—high, medium, low—advocated by TIMSS to represent mathematics and science self-concept rather than latent variables based on the continuous multiple indicators of each of these constructs. The use of these truncated (trichotomized) scale scores—even though they are provided as part of the TIMSS database—is generally unacceptable in relation to current best practice (Marsh et al., 2013). In particular, even compared to untruncated scores, this approach substantially reduces reliability, statistical power, and predictive validity (MacCallum, Zhang, Preacher, & Rucker, 2002). Additionally, the use of these scores is based on the implicit assumption that measurement error is the same across all countries and ignores method effects associated with parallel worded items and negatively worded items that are present in the TIMMS data (see the following). In contrast, appropriate latent variable models such as those used in the present study correct for measurement error and method effects and allow them to vary across countries.

Marsh et al. (2013) evaluated the factor structure of TIMSS mathematics and science constructs for eight (four Arab-speaking and four English-speaking) countries based on the eighth-grade cohort. They concluded that there was support for the a priori factor structure but that it was complicated by method effects associated with negatively worded items and parallel wording used in mathematics and science constructs. In particular, all the constructs were substantially more reliable in English-speaking countries than Arabic-speaking countries. Nevertheless, factor loadings were reasonably invariant across the eight countries. Based on their results they argued that TIMSS scale scores should not be used, that results based on scale scores are likely to be biased, and that analyses based on them—particularly those comparing findings from different countries—should be viewed with extreme caution. Instead, they argued that analyses should be based on appropriate latent variable models that controlled for measurement error and the complex factor structure. We also note that for latent variable models of differences in patterns of relations among multiple groups it is only necessary to have factor loading invariance but that studies based on manifest variables or those comparing means across the different groups require more stringent assumptions (Marsh et al., 2009; also see subsequent discussion). However, because the models considered here do not involve the comparison of latent means, we only focus on tests of the invariance of factor loadings.

Importantly, these measurement issues that undermine the Chiu (2012) study are even more critical in our developmental study in that measurement problems specific to the eighth-grade cohort are likely to be exacerbated in comparisons across the fourth- and eighth-grade samples. More specifically, an important aim of our study is to evaluate developmental hypotheses based on comparison of results from the eighth-grade cohort considered by Chiu (2012) and the fourth-grade cohort, which previously has not been evaluated in terms of the I/E model. Because these complications in TIMSS self-concept measures are so critical to tests of the I/E model we provide extensive analysis of measurement issues in the supplemental materials in the online journal.

More generally, tests of factorial invariance are a critical feature in both developmental (e.g., invariance over time or age cohorts) and cross-cultural (e.g., invariance over countries) studies. We also note that for latent variable models of differences in patterns of relations among multiple groups—including path models (e.g., Figure 1)—it is only necessary to have factor loading invariance. However, if tests were based on manifest variables, it would also be important to test the invariance of measurement errors, but we already know that such tests would fail (see Table 1). In this sense, the assumptions underlying manifest variable models are more demanding than those for latent variable models. Furthermore, if the focus was on differences in latent means across the multiple groups, it would also be important to test the invariance of intercepts (for further discussion and limitations of this approach to comparing latent means across countries, see Marsh, Hau, et al., 2006; Millsap, 2011; Nagengast & Marsh, 2013). However, because the models considered here do not involve the comparison of latent means, we only focused on tests of the invariance of factor loadings.

Table 1 Reliability of the Trends in International Mathematics and Science Study (TIMSS) Mathematics and Science Motivation Constructs Used in This Study

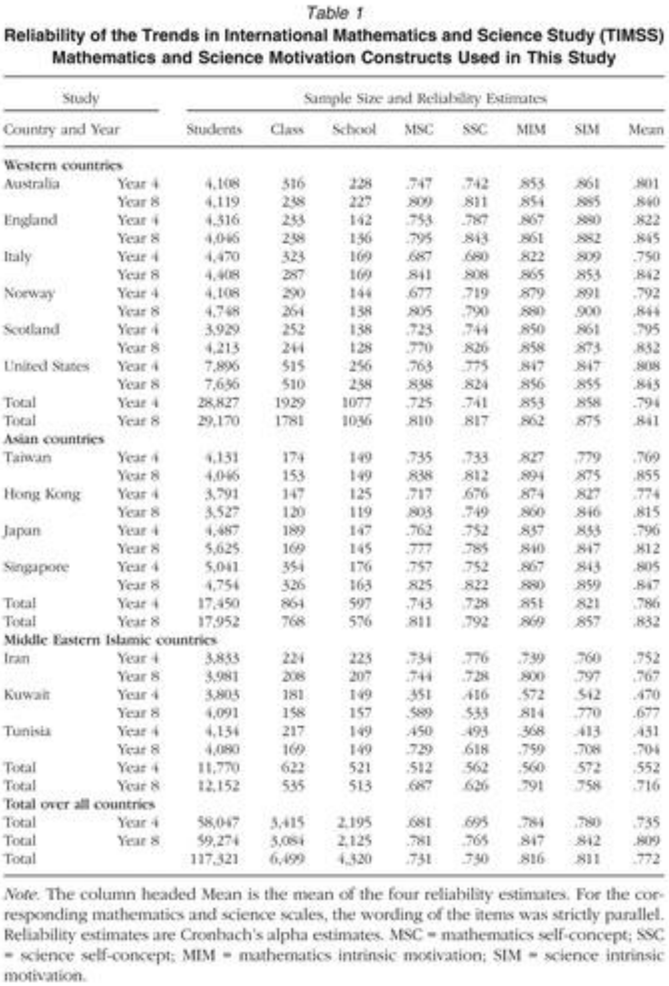

Table 1

Reliability of the Trends in International Mathematics and Science Study (TIMSS) Mathematics and Science Motivation Constructs Used in This Study

Study		Sample Size and Reliability Estimates							
Country and Year		Students	Class	School	MSC	SSC	MIM	SIM	Mean
Western countries									
Australia	Year 4	4,108	316	228	.747	.742	.853	.861	.801
	Year 8	4,119	238	227	.809	.811	.854	.885	.840
England	Year 4	4,316	233	142	.753	.787	.867	.880	.822
	Year 8	4,046	238	136	.795	.843	.861	.882	.845
Italy	Year 4	4,470	323	169	.687	.680	.822	.809	.750
	Year 8	4,408	287	169	.841	.808	.865	.853	.842
Norway	Year 4	4,108	290	144	.677	.719	.879	.891	.792
	Year 8	4,748	264	138	.805	.790	.880	.900	.844
Scotland	Year 4	3,929	252	138	.723	.744	.850	.861	.795
	Year 8	4,213	244	128	.770	.826	.858	.873	.832
United States	Year 4	7,896	515	256	.763	.775	.847	.847	.808
	Year 8	7,636	510	238	.838	.824	.856	.855	.843
Total	Year 4	28,827	1929	1077	.725	.741	.853	.858	.794
Total	Year 8	29,170	1781	1036	.810	.817	.862	.875	.841
Asian countries									
Taiwan	Year 4	4,131	174	149	.735	.733	.827	.779	.769
	Year 8	4,046	153	149	.838	.812	.894	.875	.855
Hong Kong	Year 4	3,791	147	125	.717	.676	.874	.827	.774
	Year 8	3,527	120	119	.803	.749	.860	.846	.815
Japan	Year 4	4,487	189	147	.762	.752	.837	.833	.796
	Year 8	5,625	169	145	.777	.785	.840	.847	.812
Singapore	Year 4	5,041	354	176	.757	.752	.867	.843	.805
	Year 8	4,754	326	163	.825	.822	.880	.859	.847
Total	Year 4	17,450	864	597	.743	.728	.851	.821	.786
Total	Year 8	17,952	768	576	.811	.792	.869	.857	.832
Middle Eastern Islamic countries									
Iran	Year 4	3,833	224	223	.734	.776	.739	.760	.752
	Year 8	3,981	208	207	.744	.728	.800	.797	.767
Kuwait	Year 4	3,803	181	149	.351	.416	.572	.542	.470
	Year 8	4,091	158	157	.589	.533	.814	.770	.677
Tunisia	Year 4	4,134	217	149	.450	.493	.368	.413	.431
	Year 8	4,080	169	149	.729	.618	.759	.708	.704
Total	Year 4	11,770	622	521	.512	.562	.560	.572	.552
Total	Year 8	12,152	535	513	.687	.626	.791	.758	.716
Total over all countries									
Total	Year 4	58,047	3,415	2,195	.681	.695	.784	.780	.735
Total	Year 8	59,274	3,084	2,125	.781	.765	.847	.842	.809
Total		117,321	6,499	4,320	.731	.730	.816	.811	.772

Note. The column headed Mean is the mean of the four reliability estimates. For the corresponding mathematics and science scales, the wording of the items was strictly parallel. Reliability estimates are Cronbach's alpha estimates. MSC = mathematics self-concept; SSC = science self-concept; MIM = mathematics intrinsic motivation; SIM = science intrinsic motivation.

The Present Investigation: A Priori Predictions and Research Questions

In the present study, we test the extension and generalizability of the I/E over age cohort, country, and construct, focusing on the support for the I/E model in the TIMMS data, controlling for a number of measurement concerns (Liu & Meng, 2010; Marsh et al., 2013; also see supplemental materials in the online journal for further discussion).

Support for the I/E Predictions

Extending previous research based almost completely on verbal and mathematics constructs (two domains at opposite ends of the ASC continuum) to mathematics and science (two domains close to each other on the ASC continuum), we hypothesize that overall there will be support for I/E predictions. More specifically, the effects of mathematics and science achievement will be:

- positive on matching areas of self-concept (the horizontal paths in Figure 1),
- negative on nonmatching, contrasting areas of self-concept and intrinsic motivation (the cross-paths in Figure 1).

Generalizability of Support for the I/E Predictions Over Age Cohort and Country

Constructs

We hypothesize that support for predictions will generalize over self-concept and intrinsic motivation. However, consistent with theoretical models of the formation of self-concept and previous research (e.g., Denissen et al., 2007; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005), we predict that horizontal paths will be stronger for self-concept than intrinsic motivation.

Age Cohort

Consistent with the Möller et al. (2009) meta-analysis, we hypothesize that support for predictions will generalize over age cohorts. However, in contrast to these meta-analysis results but consistent with developmental studies more generally, we hypothesize that support will be stronger for secondary than primary school students. However, as noted earlier, there are few direct tests of the generalizability of the I/E model primary and secondary age cohorts considered. Hence, this is an important focus and unique contribution of the present investigation.

Country

We hypothesize that this support will generalize over countries but will be stronger in Western and Asian countries, which have been the basis of most previous research, than in Middle Eastern Islamic countries, where there is less emphasis on evaluative and diagnostic feedback and where relations between ASC and achievement tend to be lower (see earlier discussion of research by Abu-Hilal and colleagues).

Age Cohort × Country Interactions

A unique contribution of this study is the juxtaposition of developmental hypotheses about age cohort effects and cross-cultural psychology hypotheses about country-level differences. We hypothesize that support for the I/E model will generalize over all age cohort by country combinations but leave as a research question whether there will be meaningfully large age cohort by country interactions in the sizes of either the horizontal or cross-paths (see Figure 1). In a substantive methodological synergy (Marsh & Hau, 2007), we also develop new methodological approaches to the evaluation of this interaction, capitalizing on the flexibility available in the Mplus statistical package.

New Contributions

Our study has important theoretical, developmental, cross-cultural, and methodological implications for I/E model in particular, but also for more general research on the formation of self-concept in relation to frames of reference. Developmentally, our study fills a critical gap in studies of the generalizability of the I/E model over primary and secondary age cohorts. In contrast to the comprehensive meta-analysis of I/E studies, but consistent with other ASC research, we hypothesize that support for the I/E model will be stronger in secondary than in primary school students. We hypothesize and seek to test the cross-cultural generalizability of the I/E predictions and the extent to which this cross-cultural support generalizes over different ages and different domains. In particular, ours is apparently the first study to test these predictions with the Year 4 sample from TIMSS and the first to systematically compare results from nationally representative samples of primary and secondary school students from each of a diverse sample of countries. Although there is considerable research comparing relations between ASC and achievement in Western and Asian countries, including tests of the I/E model, ours is apparently the first to expand this cross-cultural perspective to test these specific hypotheses and juxtapose Western and Asian results with those from Middle Eastern Islamic cultures. We extend tests of the I/E model to evaluate generalizability of the pattern of results to intrinsic motivation and introduce new, stronger statistical models that provide more appropriate tests of our a priori hypotheses.

Method

Participants

TIMSS 2007 (Olson, Martin, & Mullis, 2008) data are based on nationally representative samples of students from participating countries using a two-stage cluster design, sampling schools and intact classrooms from the target grade in the school (for more details, see the TIMSS 2007 Technical Report by Olson et al., 2008). For the present investigation we consider data from a total of 117,321 students in 6,499 fourth- and eighth-grade classes in six Western countries (Australia, England, Italy, Norway, Scotland, and United States), four Asian countries (Hong Kong, Japan, Singapore, and Taiwan), and three Middle Eastern Islamic countries (Iran, Kuwait, and Tunisia) where science was taught as an integrated subject and where data were available for both fourth- and eighth-grade cohorts (many countries only collect TIMSS data for eighth-grade students; see Table 1 for the number of students, classes, and schools that were sampled from each country in each age cohort.) In all countries, the materials were administered near the end of the school year (typically October or November in the Southern Hemisphere and April to June in the Northern Hemisphere).

Student achievement scores in TIMSS (Olson et al., 2008) were developed based on item response theory (IRT). For the eighth-grade tests the content domains for science were biology, chemistry, physics, and earth sciences; for mathematics they were algebra, data and chance, number, and geometry. For the fourth-grade tests the science content domains were biology, physical science, and earth science; for mathematics they were algebra, data and chance, number, geometric shapes and measures, and data display. In both subject domains and age cohorts, achievement test items involved a mixture of constructed response and multiple choice items that involved a mixture of processes (knowing, applying, and reasoning). The final items were selected on the basis of item analyses of responses from large-scale pilot studies. As noted earlier, students in both age cohorts responded to items designed to measure self-concept and intrinsic motivation in both the mathematics and science domains (see Table 2 for the wording of the items). Within each age cohort, the wording of the items for mathematics and science was strictly parallel except for the words *mathematics* or *science*. For the two age cohorts the wordings of all intrinsic motivation and two of the self-concept items were exactly the same, but there were minor wording changes for two of the self-concept items (see Table 2). Students in both age cohorts responded to all items on a classic Likert (agree-disagree) response scale.

Table 2 A Priori Factor Structure Relating Trends in International Mathematics and Science Study (TIMSS) Motivation Items to Latent Factors

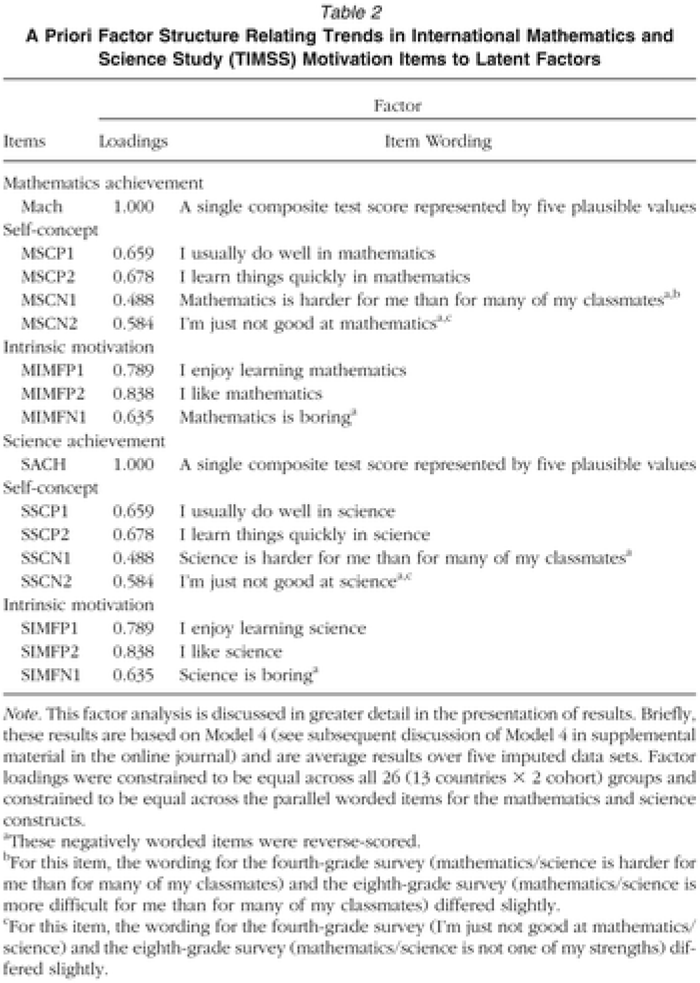

Table 2

A Priori Factor Structure Relating Trends in International Mathematics and Science Study (TIMSS) Motivation Items to Latent Factors

	Factor	
Items	Loadings	Item Wording
Mathematics achievement		
Mach	1.000	A single composite test score represented by five plausible values
Self-concept		
MSCP1	0.659	I usually do well in mathematics
MSCP2	0.678	I learn things quickly in mathematics
MSCN1	0.488	Mathematics is harder for me than for many of my classmates[a,b]
MSCN2	0.584	I'm just not good at mathematics[a,c]
Intrinsic motivation		
MIMFP1	0.789	I enjoy learning mathematics
MIMFP2	0.838	I like mathematics
MIMFN1	0.635	Mathematics is boring[a]
Science achievement		
SACH	1.000	A single composite test score represented by five plausible values
Self-concept		
SSCP1	0.659	I usually do well in science
SSCP2	0.678	I learn things quickly in science
SSCN1	0.488	Science is harder for me than for many of my classmates[a]
SSCN2	0.584	I'm just not good at science[a,c]
Intrinsic motivation		
SIMFP1	0.789	I enjoy learning science
SIMFP2	0.838	I like science
SIMFN1	0.635	Science is boring[a]

Note. This factor analysis is discussed in greater detail in the presentation of results. Briefly, these results are based on Model 4 (see subsequent discussion of Model 4 in supplemental material in the online journal) and are average results over five imputed data sets. Factor loadings were constrained to be equal across all 26 (13 countries × 2 cohort) groups and constrained to be equal across the parallel worded items for the mathematics and science constructs.

[a] These negatively worded items were reverse-scored.

[b] For this item, the wording for the fourth-grade survey (mathematics/science is harder for me than for many of my classmates) and the eighth-grade survey (mathematics/science is more difficult for me than for many of my classmates) differed slightly.

[c] For this item, the wording for the fourth-grade survey (I'm just not good at mathematics/science) and the eighth-grade survey (mathematics/science is not one of my strengths) differed slightly.

Data Analysis

Achievement test scores for each student are reported in the TIMSS 2007 database as five plausible values (Olson et al., 2008)—numbers drawn randomly from the distribution of scores that could be reasonably assigned to each student. Implementation of this approach by TIMSS was due in part to the use of a matrix sampling approach in which each student was administered only a sample of the achievement test items that were combined to form a total score using an IRT approach to test equating (Olson et al., 2008). Following recommendations by TIMSS, all data analyses with achievement were run separately for each of the five plausible values, and the results were aggregated appropriately in order to obtain unbiased estimates. Although the amount of missing data was relatively small, we used full information maximum likelihood estimation to control for missing data, noting that this was done separately for each of the five data sets based on different plausible values, and then combined using the Rubin (1987; Schafer, 1997) strategy, which was implemented automatically in Mplus. Thus, results are based on an appropriate aggregation of results across the multiple data sets to obtain appropriate parameter estimates, standard errors, and goodness-of-fit statistics. We note, however, that this strategy was used primarily to incorporate the multiple plausible values, as the amount of missing data was so small (an average of less than 2% for the rating items, none at all for the test scores).

All analyses were based on TIMSS's HOUWGT weighting variable, which incorporates six components: three having to do with sampling of the school, class, and student and adjustment factors associated with nonparticipation at the level of school, class, and student. The weighting is based on the actual number of students in each country that is appropriate for correct computation of standard errors and tests of statistical significance. Correcting for the clustering inherent in the two-stage clustering sample, the 26 (13 countries × 2 age cohorts) groups were treated as grouping variables that were the basis of the multigroup analyses, whereas the class and school clustering variables were used to control for the clustered sample (using the complex design option and robust maximum likelihood options in Mplus; Muthén & Muthén, 2008–2011). We note that the classroom is the critical clustering variable for TIMSS data because class was the sampling unit used in the TIMSS sampling design, which was based on sampling all students within intact classes; most schools are represented by a single class, and a given class might not be representative of the school from which it came.

In the present investigation we used a common metric standardization strategy, first standardizing individual indicators (rating items and test scores) in relation to the grand total sample mean and standard deviation. We then used slightly different strategies for item ratings and test scores. For the self-concept and intrinsic motivation rating items, the means and standard deviations were based on the total sample, including both age cohorts. This was facilitated by the fact that the items were largely the same for all students in each age cohort. For latent factors based on these rating items, the latent factors were then standardized in relation to a common pooled within-group standard deviation. In this respect, responses by all students were placed on a common metric designed to facilitate interpretations. For the science and mathematics achievement test scores, substantially different sets of test items were needed to evaluate achievement in the two age cohorts; not even the number of content areas assessed in each age cohort was the same. Hence, we standardized the test scores separately for each age cohort using a common pooled within-group standard deviation across countries within each age cohort (for further discussion of standardization issues, see Appendix 3 in the online journal).

Estimation

Analyses conducted with Mplus 6.1 (Muthén & Muthén, 2008–2011) consisted of confirmatory factor analysis (CFA) and structural equation models (SEMs) based on the Mplus robust maximum likelihood estimator (MLR), with standard errors and tests of fit that were robust in relation to non-normality of observations and the use of categorical variables where there were at least four or more response categories, particularly when non-normality was not excessive and a design-based correction (Mplus's complex design option) was used to control for the non-independence of observations (Muthén & Muthén, 2008–2011). In the decomposition of group (13 countries × 2 age cohorts) into variance components and more detailed factorial (ANOVA-like) contrasts, we relied heavily on the flexibility of the "model constraint" function in Mplus and the resulting tests of statistical significance based on these model constraints. Thus, for example, we used these constraints to obtain ANOVA-like estimates of the proportion of variation in the horizontal and cross-paths posited in the I/E model, which were explained by the 13 countries (and three groups of countries: Western, Asian, Middle Eastern Islamic), two age cohorts (Grade 4 vs. 8), and age cohort by country interactions. These were followed by more specific tests of a priori hypotheses. This evolving methodology—combining the flexibility typically associated with analyses of manifest variables with latent variable models—is apparently a new methodological contribution of the present investigation with broad applicability to cross-cultural and educational research more generally.

Preliminary Analyses

Before evaluating support for our predictions, it is critical to identify, evaluate, and control for a number of measurement issues associated with the TIMSS database. Marsh et al. (2013) provide an extensive critique of the use of manifest trichotimization of psychological scales scores, as reported in the TIMSS manual. They suggest that this use affects power and reliability. Thus, the high standards of the achievement tests developed by TIMSS are not reflected in the student survey. Marsh et al. further suggest that the use of trichotimized scales scores is particularly problematic for TIMSS, as reliabilities vary systematically from country to country (also see Appendix 1 in the online journal) and there are clear method effects present: (1) parallel wording for items used to infer mathematics and science constructs (e.g., “I usually do well in mathematics” and “I usually do well in science”) and (2) a mixture of positively and negatively worded items within the same construct (e.g., “I usually do well in mathematics” and “I am just not good in mathematics”). Following recommendations to overcome these problems (Marsh et al., 2013), we utilize latent variable models to model both parallel items and account for negative wording of items. In addition, comparison of results across different countries, age cohorts, or content domains (i.e., mathematics and science) requires strong assumptions about the invariance of the factor structure. If the underlying factors differ fundamentally in different groups, then there is no basis for interpreting observed differences (the “apples and oranges” problem; see Millsap, 2011). Marsh et al. (2013) provide an extensive exploration of these issues in TIMSS in relation to a sample of countries based on the TIMSS eighth-grade cohort but did not consider the TIMSS fourth-grade cohort or as extensive a range of countries as considered here. Thus, in preliminary analyses we considered these issues; the results are presented in detail in the supplemental materials in the online journal. In summary, the results suggest:

1. Reliability estimates were systematically higher for the older age cohort and systematically lower in Middle Eastern Islamic countries than Western or Asian countries (Table 1). Reliability estimates on average were acceptable but in some cases were not suitable for manifest models, as typically are conducted in TIMSS research. To address this problem, we utilized latent variable models that control for measurement error (Millsap, 2011).
2. Method effects associated with parallel and negatively worded items were explored in a series of CFA models, following from previous work by Marsh et al (2013). Model fit (see Appendix 1 in the online journal for a discussion of goodness of fit) suggests that both sources of method effect contributed substantially and independently to goodness of fit. As a result, all models in the following section contained correlated residuals that accounted for method effects relating to parallel and negative item wording.
3. Invariance of factor structure is a critical assumption of cross-cultural research (see Marsh et al., 2013; Parker, Dowson, & McInerney, 2007). We found reasonable support for the invariance of factor loadings over construct, age cohort, and country. All analyses reported in the following are based on full factor loading invariance.

Results

In the evaluation of support for I/E predictions, we focus on horizontal paths predicted to be positive, cross-paths predicted to be negative (see Figure 1), and a priori predictions about generalizability in relation to the two constructs (self-concept vs. intrinsic motivation), the two age cohorts, and the 13 different countries (5 Western, 4 Asian, and 3 Middle Eastern Islamic). All 208 paths (4 paths × 2 constructs × 2 age cohorts × 13 countries) are presented in Appendix 2 in the online journal, along with standard errors and tests of significance for effects of country, age cohort, and their interaction. However, to facilitate summary and discussion of the results, in Table 3 we have also computed the mean of the horizontal and cross-paths for each of the 26 (13 countries × 2 age cohort) groups, along with *SE*s and tests of statistical significance. Because of the large sample sizes, even small differences are statistically significant. Thus, our focus is on the sizes of the effects (represented by standardized path coefficients).

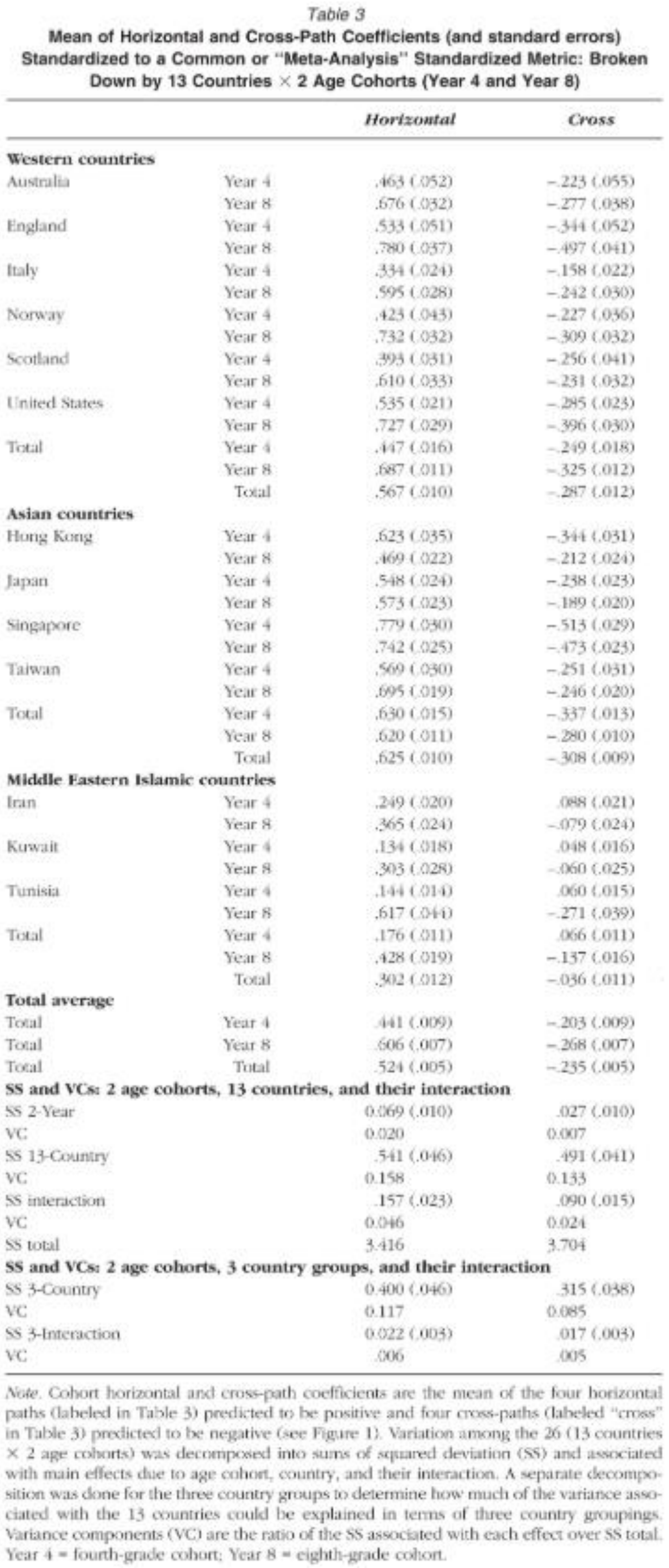

Table 3

Mean of Horizontal and Cross-Path Coefficients (and standard errors) Standardized to a Common or "Meta-Analysis" Standardized Metric: Broken Down by 13 Countries × 2 Age Cohorts (Year 4 and Year 8)

		Horizontal	*Cross*
Western countries			
Australia	Year 4	.463 (.052)	–.223 (.055)
	Year 8	.676 (.032)	–.277 (.038)
England	Year 4	.533 (.051)	–.344 (.052)
	Year 8	.780 (.037)	–.497 (.041)
Italy	Year 4	.334 (.024)	–.158 (.022)
	Year 8	.595 (.028)	–.242 (.030)
Norway	Year 4	.423 (.043)	–.227 (.036)
	Year 8	.732 (.032)	–.309 (.032)
Scotland	Year 4	.393 (.031)	–.256 (.041)
	Year 8	.610 (.033)	–.231 (.032)
United States	Year 4	.535 (.021)	–.285 (.023)
	Year 8	.727 (.029)	–.396 (.030)
Total	Year 4	.447 (.016)	–.249 (.018)
	Year 8	.687 (.011)	–.325 (.012)
	Total	.567 (.010)	–.287 (.012)
Asian countries			
Hong Kong	Year 4	.623 (.035)	–.344 (.031)
	Year 8	.469 (.022)	–.212 (.024)
Japan	Year 4	.548 (.024)	–.238 (.023)
	Year 8	.573 (.023)	–.189 (.020)
Singapore	Year 4	.779 (.030)	–.513 (.029)
	Year 8	.742 (.025)	–.473 (.023)
Taiwan	Year 4	.569 (.030)	–.251 (.031)
	Year 8	.695 (.019)	–.246 (.020)
Total	Year 4	.630 (.015)	–.337 (.013)
	Year 8	.620 (.011)	–.280 (.010)
	Total	.625 (.010)	–.308 (.009)
Middle Eastern Islamic countries			
Iran	Year 4	.249 (.020)	.088 (.021)
	Year 8	.365 (.024)	–.079 (.024)
Kuwait	Year 4	.134 (.018)	.048 (.016)
	Year 8	.303 (.028)	–.060 (.025)
Tunisia	Year 4	.144 (.014)	.060 (.015)
	Year 8	.617 (.044)	–.271 (.039)
Total	Year 4	.176 (.011)	.066 (.011)
	Year 8	.428 (.019)	–.137 (.016)
	Total	.302 (.012)	–.036 (.011)
Total average			
Total	Year 4	.441 (.009)	–.203 (.009)
Total	Year 8	.606 (.007)	–.268 (.007)
Total	Total	.524 (.005)	–.235 (.005)
SS and VCs: 2 age cohorts, 13 countries, and their interaction			
SS 2-Year		0.069 (.010)	.027 (.010)
VC		0.020	0.007
SS 13-Country		.541 (.046)	.491 (.041)
VC		0.158	0.133
SS interaction		.157 (.023)	.090 (.015)
VC		0.046	0.024
SS total		3.416	3.704
SS and VCs: 2 age cohorts, 3 country groups, and their interaction			
SS 3-Country		0.400 (.046)	.315 (.038)
VC		0.117	0.085
SS 3-Interaction		0.022 (.003)	.017 (.003)
VC		.006	.005

Note. Cohort horizontal and cross-path coefficients are the mean of the four horizontal paths (labeled in Table 3) predicted to be positive and four cross-paths (labeled "cross" in Table 3) predicted to be negative (see Figure 1). Variation among the 26 (13 countries × 2 age cohorts) was decomposed into sums of squared deviation (SS) and associated with main effects due to age cohort, country, and their interaction. A separate decomposition was done for the three country groups to determine how much of the variance associated with the 13 countries could be explained in terms of three country groupings. Variance components (VC) are the ratio of the SS associated with each effect over SS total. Year 4 = fourth-grade cohort; Year 8 = eighth-grade cohort.

Support for I/E Predictions

Overall there was good support for predictions (see Table 3). Averaged across all groups and constructs, horizontal paths were significantly positive (.524, *SE* = .005, Table 3), while cross-paths were significantly negative (−.235, *SE* = .005). There were substantial country-level differences in both the horizontal and cross-paths, and these country differences interact with age cohort (Table 3). Particularly noticeable are differences in the Middle Eastern Islamic countries in that predictions are not fully supported for the youngest cohort. Although the horizontal paths were significantly positive, they were substantially smaller in Middle Eastern Islamic countries (.176, *SE* = .011) than in the Western (.447, *SE* = .016) and particularly in the Asian (.630, *SE* = .016) countries. However, the cross-paths for the youngest cohort of Middle Eastern Islamic students were slightly (significantly) positive (.066, *SE* = .011) rather than negative. Nevertheless, support for predictions is clearly evident for the older cohort of Middle Eastern Islamic students, even though, compared to other countries, these horizontal paths were still less positive (Middle Eastern Islamic: .438, *SE*= .019; Western: .687, *SE* = .011; Asian: .620, *SE* = .011), while the cross-paths were less negative (Middle Eastern Islamic: −.137, *SE* = .016; Western: −.325, *SE* = .012; Asian: −.280, *SE* = .009). Consistent with a priori predictions, this overall support for the I/E model was significantly stronger for the older cohort; horizontal paths were significantly more positive (.606 vs. .441, $p < .001$; see Table 3), and cross-paths were significantly more negative (−.268 vs. −.203, $p < .01$).

In order to reduce the complexity of the presentation, we have only presented results averaged over constructs (math and verbal responses to ASC and intrinsic motivation). However, the pattern of results considered separately for each construct (see Appendix 2 in the online journal) is consistent with those presented here for the averaged results. In particular, horizontal paths are significantly positive, whereas cross-paths are significantly negative for all but the Year 4 cohort of Middle Eastern Islamic students. Although the pattern of support for the I/E model generalizes across constructs, consistent with a priori predictions, the horizontal paths in particular were systematically stronger for self-concept ratings (.663 and .564) than for intrinsic motivation ratings (.440 and .426). Interestingly, however, the negative cross-paths were as large or larger for intrinsic motivation (−.218 and −.317) as for self-concept (−.194 and −.251). Nevertheless, the pattern of country and country by age cohort differences was consistent over the two constructs. In particular, lack of support for I/E predictions for young cohorts in Middle Eastern Islamic countries (non-negative cross-paths) was evident for both self-concept and intrinsic motivation.

Discussion

The present investigation—along with the Marsh and Hau (2004) PISA study and Möller et al.'s (2009) meta-analysis—provides the strongest support for the generalizability of both external and internal frame of reference effects posited in the I/E model. However, the present study has important advantages over previous studies.

Importance of Latent Variable Models

The importance of using more appropriate latent variable models in the present investigation rather than the TIMSS scale scores that have been used in most TIMSS studies is highlighted by comparing our eighth-grade results using latent variable models with those based on the Chiu (2012) analysis of TIMSS2003 data using the TIMSS scale scores. The horizontal paths in our study are systematically more positive (.765, *SE* = .011 vs. .55, *SE* = .03, mathematics achievement to science self-concept; .659, *SE* = .014 vs. .40 *SE* = .03, science achievement to MSC). These higher values are consistent with the control for measurement error in the latent variable models. The cross-paths in our study are more negative (−.222, *SE* = .011 vs. −.09, *SE* = .03, mathematics achievement to MSC; −.351, *SE* = .012 vs. −.10, *SE* = .03, science achievement to science self-concept). Furthermore, because the reliability estimates vary substantially for different countries, the extent of bias also varies substantially for different countries. These differences are consistent with our claim that analyses based on TIMSS scale scores should not be used (see earlier discussion and supplemental materials) and that results based on them are likely to be biased and should be viewed with extreme caution. In this respect, the present investigation provides strong support for substantive-methodological synergies (Marsh & Hau, 2007) in which complex substantive issues with important theoretical and policy/practice implications for applied educational research typically require the application of new and evolving statistical methodology.

Developmental Perspectives

Of particular relevance, nearly all I/E studies—and particularly the cross-national studies—have been based on responses by secondary students. Almost no research—and particularly no cross-cultural research—has been done with primary students, although clearly this is important. In particular there have been no comparisons between matched, nationally representative age cohorts of primary and secondary students. While a few studies of primary students were included in Möller et al.'s (2009) meta-analysis, the developmental aspect of their study was not systematically evaluated— due in part to the paucity of available data on young children. Furthermore, because previous research has not been based on matched, nationally representative samples of students of different ages, the samples included in the meta-analyses are not directly comparable. In this respect, the TIMSS 2007 data are ideally suited to evaluating the juxtaposition of nationality, age cohort, and their interaction. Coupled with new statistical models exploiting the flexibility of Mplus (see Appendix 4 in the online journal), we systematically evaluated main effects and interaction effects based on latent path coefficients and pursued detailed comparisons of the interaction effects. This methodological-substantive synergy was important, providing more appropriate tests of developmental perspectives and their cross-cultural generalizability that are important contributions of the present investigation. In particular, not only did we provide apparently the first results showing that support for I/E predictions were stronger for secondary students than primary students (more positive effects for horizontal paths, more negative effects for cross-paths), but we also showed that these developmental differences varied as a function of country in ways consistent with a priori predictions.

Western, Asian, and Middle Eastern Islamic Cultures

Although the focus of the cross-cultural component of our study from a universalist perspective of the generalizability support across countries, our study is apparently the first to specifically compare I/E results in a sample of Middle Eastern Islamic countries with those from Asian and Western countries, which have been the basis of most I/E studies (see Möller et al., 2009, meta-analysis). Consistent with a priori predictions based on research by Abu-Hilal and colleagues (see earlier discussion), support for the I/E model was systematically weaker in the Middle Eastern Islamic countries, particularly for the younger age cohort. This prediction was based on previous research showing that Middle Eastern Islamic students do not receive as much evaluative feedback about their achievement as do Western and Asian students and are not socialized in such a way as to critically evaluate their academic skills in relation to classmates. Hence, not only were cross-paths close to zero for the fourth-grade cohort of Middle Eastern Islamic students, but the effect of achievement on self-concepts in matching academic areas (the horizontal paths in the I/E model) was also substantially lower than for Western and Asian students. Indeed, consistent with speculations by Abu-Hilal and Bahri (2000) that the relation of ASC formation with achievement in Middle Eastern Islamic middle school students was similar to that found in younger students from Western countries, support for the I/E model for eighth-grade Middle Eastern Islamic students was similar to that found for the fourth-grade cohort in the Western countries (for further discussion, see Abu-Hilal, 2001; Abu-Hilal & Aal-Hussain, 1997; Abu-Hilal & Bahri, 2000). Although this was not predicted a priori, it was interesting to observe that self-concepts of the fourth-grade cohort were more strongly associated with matching areas of achievement (the horizontal paths in the I/E model) for Asian students than Western students. While beyond the scope of the present investigation, we speculate that young Asian students receive more diagnostic, evaluative feedback and are socialized to compare their academic accomplishments with classmates more than are young Western students, leading in part to stronger relations between academic achievement and self-concept. Indeed, an important direction for further research would be to take a more emic or case study approach into a more nuanced understanding of the differences and similarities between cultures, cultural norms, approaches to teaching, and learning within each of the countries and how these may influence support for the I/E model.

Limitations and Directions for Further Research

Standardization Issues

Inherent in the use of academic achievement measures for students of different ages is the question of the comparability of the achievement tests for the different age cohorts. Thus, for example, the TIMSS achievement tests for the fourth- and eighth-grade cohorts are based on completely different sets of items designed to measure somewhat different content areas. Here we used a common metric standardization approach, greatly facilitated by the fact that we had matched nationally representative samples for each age cohort for all countries considered here. Nevertheless, we found the surprising result that variability in achievement test scores from the same country was not consistent across the two age cohorts. In response to this issue, we also pursued supplemental analyses based on within-group standardization approach (i.e., standardization separately within each of the 26 age/country groups) like that typically used in meta-analysis (e.g., Möller et al., 2009). Broadly the pattern of results was similar for both standardizations, but the meta-analysis standardization resulted in somewhat different interpretations, particularly for Asian countries (Appendix 3 in the online journal). In particular, support for the prediction that path coefficients in support of the I/E model (positive horizontal, negative cross-paths) would be larger for the older eighth-grade cohort than the fourth-grade cohort is stronger with the meta-analysis standardization. These differences are easily explained in terms of *SD*s of achievement tests for the different countries (*SD*s in the fourth- and eighth-grade cohorts were similar in Western countries, substantially larger for the older cohort in Asian countries, and substantially larger for the younger cohort in in the Middle Eastern Islamic countries) such that support for I/E predictions was confounded by these age differences in Asian countries. Although it is unclear whether this represents problems in the scaling of the scores or in our standardization strategies, or substantively important differences, the results warrant further investigation. The comparability of scores across age cohorts for self-concept and intrinsic motivation was facilitated by the use of the same items (with minor exceptions), coupled with good support for the invariance of factor loadings across age cohorts. Indeed, although there was reasonable support for the complete invariance of factor loadings across both country and age cohort, the support for invariance was actually stronger across age cohort within countries than across countries within age cohorts. Nevertheless, the substantial age cohort and country differences in the extent of measurement error and method effects dictate caution in the interpretation of these results.

Extending Dimensional Comparison Theory

The I/E model is based on the assumption that students engage in internal dimensional comparisons—as well as social comparison—in the formation of their self-concepts in different school subjects. Implicit in the prediction of negative cross-paths is the assumption that such dimensional comparisons result in contrast effects, such that the better I am in one subject the lower my self-concept is in a contrasting school subject. As suggested by Möller and Marsh (2013; also see Marsh, 1986; Marsh et al., in press; Möller, Streblow, & Pohlmann, 2006; Xu et al., 2013), this implies that students believe that there is a negative interdependence between the two subjects. Hence, most of the extensive support for I/E predictions is based primarily on self-concept responses by secondary students to mathematics and verbal domains, which represent opposite ends of the theoretical continuum of ASCs. However, the theoretical underpinning of the I/E model posits that the negative contrast effects will diminish and might even become positive (assimilation) when dimensional comparisons are based on domains that are close together on the academic continuum. Thus, if students consider abilities in two subjects to be mutually supportive, the paths from achievement in one domain to self-concept in a closely related subject should be less negative or even positive—an assimilation effect. For example, Möller et al. (2006) found that mathematics achievement had a positive effect on physics self-concepts. Similarly, recent research (Parker, Marsh, et al., 2013) based on PISA data indicates that paths leading to MSC (the only self-concept domain assessed) were positive for mathematics achievement and negative for verbal achievement. Critically, the path from science achievement (a domain close to mathematics on the underlying academic continuum) to MSC was also positive; science achievement was positively predictive of MSC, even after controlling for science (and verbal) achievement: an assimilation effect. Nevertheless, for both studies of social and dimensional comparison, there is consistent support for contrast effects, whereas assimilation effects—a priori or post hoc—have been elusive (e.g., Marsh et al., 2008; Möller et al., 2009). An interesting direction for further research is to explore more fully individual student differences, domains, or conditions under which cross-paths are positive rather than negative (i.e., where there is assimilation rather than contrast).

Implications for Practice

The I/E model also has practical implications for educational practice and understanding of self-concept formation. This theoretical model is one of the dominant models of self-concept formation in educational research with surprising—even paradoxical—results about how being good in one domain undermines self-concept in another domain. The model has fundamental implications about the way teachers give feedback to students and understand their students' self-perceptions of their relative strengths/weaknesses in different domains. However, the vast majority of this research is based on secondary students from Western countries. There has been almost no good developmental research into how support for the model generalizes across primary and secondary students and almost no tests of the model outside of Western and a few Asian countries. In this respect, the present investigation has profound practical implications for the way teachers understand and relate to their students.

When teachers, parents, and other significant others are asked to infer the students' ASCs (see Dai, 2002; Marsh, 2007), their responses reflected primarily the external comparison process, so that inferences were not nearly so domain specific as responses by students. Their responses imply that students who are bright in one area tend to be seen as having good ASCs in all areas (consistent with corresponding measures of achievement), whereas students who are not bright in one area are seen as having poor ASC in all areas. However, if teachers and significant others better understood formation of self-concepts in different academic domains, they would be able to better understand their students and provide more appropriate feedback that is credible, particularly for less able students. Even bright students might have an average or below average self-concept in their weakest school subjects, which may seem paradoxical in relation to their good achievement (good relative to other students but not relative to their own performance in other school subjects). Similarly, even poor students may have an average or above average self-concept in their best school subject that may seem paradoxical in relation to their below-average achievement in that subject (but not relative to their other school subjects). This is especially important for primary school students who were a major focus of the present investigation, in that primary school teachers generally instruct students across multiple school subjects, are more likely to be called on to evaluate student noncognitive outcomes as part of their reporting of progress to parents, and are expected to provide a nurturing role in developing student self-perceptions as a confident learner in different school subjects. Indeed, when policy statements and polic makers refer to accountability issues, they typically refer to standardized achievement test scores. However, many self-concept studies (see Marsh, 2007) show that ASC is sometimes a more important determinant of critical educational choices than achievement. Hence, teachers and policymakers need to understand these complex issues in the way ASCs are formed.

Notes

This paper was made possible in part by a grant from the Australian Research Council (DP130102713) awarded to the first, third and fourth authors. The authors would like to thank Tihomir Asparouhov, Matthias Von Davier, Anna Preuschoff, and Michael Martin for helpful comments at earlier stages of this research.

We note that all the Islamic Middle Eastern countries in our sample are officially self-proclaimed as Islamic countries and that this distinguishes them from non-Islamic Middle Eastern countries. Although we do not have access to religious affiliation at the student level, most of the students in these countries are Islamic and would certainly be more homogeneous in relation to religion than other countries. Thus, the Muslim percentage of the total population (Pew-Templeton, 2012) in the three countries considered here are: Iran (99.6%), Kuwait (86.4%), and Tunisia (99.8%).

References

Abu-Hilal, 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. 205–230). 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. doi:10.1177/ 0022022197285002

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 (in Arabic).

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

Albert, S. (1977). Temporal comparison theory. Psychological Review, 84, 485–503. doi:10.1037/0033-295X.84.6.485

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

Bong, M. (1998). Tests of the internal/external frames of reference model with subject-specific academic self-efficacy and frame-specific academic self-concepts. Journal of Educational Psychology, 90, 102–110. doi:10.1037/0022-0663.90.1.102

Chapman, J. W., & Tunmer, W. E. (1995). Development of children's reading self-concepts: An examination of emerging subcomponents and their relation with reading achievement. Journal of Educational Psychology, 87, 154–167.

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. (2012). The internal/external frame of reference model, big-fish-littlepond effect, and combined model for mathematics and science. Journal of Educational Psychology, 104, 87–107. doi:10.1037/a0025734

Dai, D. Y. (2002). Incorporating parent perceptions: A replication and extension study of the internal/external reference model of self-concept development. Journal of Adolescent Research, 17, 617–664. doi: 10.1177/074355802237467

Davis-Kean, P. E., & Sandler, H. M. (2001). A meta-analysis of measures of self-esteem for young children: A framework for future measures, Child Development, 72, 887–906.

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

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I'm able, and I know I am: Longitudinal couplings between domain-specific achievement, self concept,and interest. Child Development, 78(2), 430–447. doi:10.1111/j.1467-8624.2007.01007.x

Dickha¨user, O. (2003). U¨ berpru¨ fung des erweiterten Modells des internal/external frame of reference [Test of the extension of the internal/external frame of reference model]. Zeitschrift fu¨r Entwicklungspsychologie und Pa¨dagogische Psychologie, 35, 200–207.

Diener, E. (2000). Subjective well-being—The science of happiness and a proposal for a national index. American Psychologist, 55(1), 34–43. doi:10.1037/0003-066X.55.1.34

Eccles, J. S. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist, 44, 78–89. doi:10.1080/00461520902832368

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

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. Personality and Social Psychology Bulletin, 21(3), 215–225. doi:10.1177/0146167295213003

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

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

Eder, R. A., & Mangelsdorf, S. C. (1997). The emotional basis of early personality development: Implications for the emergent self-concept. In R. Hogan, J. Johnson & S. Briggs (Eds.), Handbook of personality psychology (pp. 209–240). San Diego, CA: Academic Press.

Goetz, T., Frenzel, A. C., Pekrun, R., & Hall, N. C. (2006): The domain specificity of academic emotional experiences. The Journal of Experimental Education, 75, 5–29.

Goetz, T., Frenzel, C. A., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. Contemporary Educational Psychology, 33, 9–33. doi:10.1016/ j.cedpsych.2006.12.002

Harter, S. (1999). The construction of the self: A developmental perspective. New York, NY: Guilford Press

Harter, S. (2006). The self. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), Handbook of child psychology: Vol. 3. Social, emotional, and personality development (6th ed., pp. 505–570). Hoboken, NJ: John Wiley.

Harter, S. (2012). The construction of the self: Developmental and sociocultural foundations (2nd ed.). New York, NY: Guilford Press.

Hattie, J. (1992). Self-concept. Mahwah, NJ: Lawrence Erlbaum associates, Inc.

Hattie, J., & Marsh, H. W. (1996). Future directions in self-concept research. In B. A. Bracken (Ed.), Handbook of self-concept (pp. 421–462). Oxford, UK: John Wiley & Sons.

James, W. (1963). The principles of psychology. New York: Holt, Rinehart & Winston. (Original work published 1890)

Kagen, S. L., Moore, E., & Bredekamp, S. (1995). Considering children's early development and learning: Toward common views and vocabulary (Report N. 95-03). Washington, DC: National Education Goals Panel.

Kagitcibasi, C., & Poortinga, Y. H. (2000). Cross-cultural psychology: Issues and overarching themes. Journal of Cross-Cultural Psychology, 31(1), 129–147.

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.

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

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. American Educational Research Journal, 23, 129–149.

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

Marsh, H. W. (1990a). The influences of internal and external frames of reference on the formation of English and math self-concepts. Journal of Educational Psychology, 82, 107–116.

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

Marsh, H. W. (1994). Using the National Educational Longitudinal Study of 1988 to evaluate theoretical models of self-concept: The Self-Description Questionnaire. Journal of Educational Psychology, 86, 439–456.

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. C., & Parker, P. (2013). Factorial, convergent, and discriminant 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., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science, 1, 133–163. doi:10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Craven, R. G., & Debus, R. (1998). Structure, stability, and development of young children’s self-concepts: A multicohort-multioccasion study. Child Development, 69(4), 1030–1053.

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., Ellis, L. A., & Craven, R. G. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. Developmental Psychology, 38, 376–393. doi: 10.1037/0012-1649.38.3.376

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

Marsh, H. W., & Hau, K-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. Contemporary Educational Psychology, 32, 151–171. doi:10.1016/j.ced psych.2006.10.008

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

Marsh, H. W., Kong, C. K., & Hau, K. (2001). Extension of the internal/external frame of reference model of self-concept formation: Importance of native and nonnative languages for Chinese students. Journal of Educational Psychology, 93, 543–553. doi:10.1037/0022-0663.93.3.543

Marsh, H. W., Martin, A. J., & Hau, K-T. (2006). A multiple method perspective on self-concept research in educational psychology: A construct validity approach. In M. Eid & E. Diener (Eds.), Handbook of multimethod measurement in psychology (pp. 441–456). Washington, DC: American Psychological Association:.

Marsh, H. W., Mo¨ller, J., Parker, P., Xu, M. K., Benjamin, N., & Pekrun, R. (in press). Internal/external frame of reference model. In J. D. Wright (Ed.), The international encyclopedia of the social and behavioral sciences (2nd ed.). Oxford, UK: Elsevier.

Marsh, H. W., Muthe´n, B., Asparouhov, T., Lu¨dtke, 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(3), 439–476. doi:10.1080/10705510903008220

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

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. Educational Psychologist, 20(3), 107–123. doi:10.1207/s1532698 5ep2003_1

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

Matsumoto, D. (2001). Cross-cultural psychology in the 21st century. In J. S. Halonen & S. F. Davis (Eds.), The many faces of psychological research in the 21st century. Retrieved from http://teachpsych.lemoyne.edu/teachpsych/faces/script/

ch05.htm

Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York, NY: Routledge/Taylor & Francis Group.

Mo¨ller, J. (2005). ‘‘Paradoxical’’ effects of praise and criticism: Social, dimensional, and temporal comparisons. British Journal of Educational Psychology, 75, 275–295. doi:10.1348/000709904X24744

Mo¨ller, J., & Ko¨ller, O. (2001). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. Journal of Educational Psychology, 93, 826–835. doi:10.1037/0022-0663.93.4.826

Mo¨ller, J., & Marsh, H. W. (2013). Dimensional comparison theory. Psychological Review, 120(3), 544–560. doi: 10.1037/a0032459

Mo¨ller, J., Pohlmann, B., Ko¨ller, O., & Marsh, H.W. (2009). A meta-analytic path analysis of the Internal/External frame of reference model of academic achievement and academic self-concept. Review of Educational Research, 79, 1129–1167. doi:10.3102/0034654309337522

Mo¨ller, J., Retelsdorf, J., Ko¨ller, O., & Marsh, H. W. (2011). The reciprocal I/E model: An integration of models of relations between academic achievement and selfconcept. American Educational Research Journal, 48, 1315–1346. doi:10.3102 /0002831211419649

Mo¨ller, J., Streblow, L., & Pohlmann, B. (2006). The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. British Journal of Educational Psychology, 76, 57–70. doi:10.1348/ 000709905X37451

Mo¨ller, J., Streblow, L., Pohlmann, B., & Ko¨ller, O. (2006). An extension to the internal/external frame of reference model to two verbal and numerical domains. European Journal of Psychology of Education, 21(4), 467–487. doi:10.1007/ BF03173515

Muthe´n, L. K., & Muthe´n, B. O. (2008–2011). Mplus user’s guide. Los Angeles, CA: Muthe´n & Muthe´n.

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup SEMs across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), A handbook of international large-scale assessment data analysis (pp. 317–344). London: Chapman & Hall.

Nagy, G., Trautwein, U., Baumert, J., Ko¨ller, O., & Garrett, J. (2006). Gender and course selection in upper secondary education: Effects of academic self-concept and intrinsic value. Educational Research and Evaluation, 12, 323–345. doi:10.1080/13803610600765687

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

Parker, P., Dowson, M., & McInerney, D. (2007). Standards for quantitative research in diverse sociocultural contexts. In D. McInerney, S. Van Etten, & M. Dowson (Eds.). Standards in education (pp. 315–330). Charlotte, NC: Information Age Press.

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S., & Abduljabbar, A. S. (2013). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. Educational Psychology. Retreived from http:// www.tandfonline.com/doi/abs/10.1080/01443410.2013.797339#preview. doi: 10.1080/01443410.2013.797339

Parker, P. D., Marsh, H. W., Lu¨dtke, O., & Trautwein, U. (2013). Differential school contextual effects for math and English: Integrating the big-fish-little-pond effect and the internal/external frame of reference. Learning and Instruction, 23, 78–89. http://dx.doi.org/10.1016/j.learninstruc.2012.07.001

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012). Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study. Developmental Psychology, 48(6), 1629–1642. doi:10.1037/a0029167

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18, 315–341. doi: 10.1007/s10648-006-9029-9

Pew-Templeton. (2012). Table: Muslim population by country. Retrieved from http:// features.pewforum.org/muslim-population/.

Renninger, K. (2009). Interest and identity development in instruction: An inductive model. Educational Psychologist, 44(2), 105–118. doi:10.1080/004615 20902832392

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, NY: Wiley.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. New York, NY:

Chapman and Hall.

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. Journal of Personality and Social Psychology, 58(5), 878–891.

Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioral research. American Psychologist, 53, 1101–1110.

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

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

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. Review of Educational Research, 46, 407–444.

Skaalvik, E. M., & Hagtvet, K. A. (1990). Academic achievement and self-concept: An analysis of causal predominance in a developmental perspective. Journal of Personality and Social Psychology, 58, 292–307. doi: 10.1037/0022-3514.58.2.292

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

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. Developmental Review, 12, 265–310.

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

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

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

Xu, M. K., Marsh, H. W., Hau, K. T., Ho, I. T., Morin, A. J., & Abduljabbar, A. S. (2013). The internal/external frame of reference of academic self-concept: Extension to a foreign language and the role of language of instruction. Journal of Educational Psychology, 105, 489–503.

Yeung, A. S., Lee, J. C., & Wong, H. (2001, April). Testing Marsh's (1986) frame of reference model of self-concept with bilingual students. Paper presented at the AERA Annual Meeting, New Orleans, LA.