Supplemental Appendix 2
TIMSS2007 Factor Structure and Measurement Invariance: Background to the Present Investigation

Liu and Meng (2010) examined the factor structure of attitudinal items from TIMSS2003. Their particular focus was on the comparison of self-concept responses by 8th grade students in three East Asian countries and the US. They were critical of researchers who used these items without appropriate recognition of the underlying constructs represented by the items. They conducted exploratory factor analyses of responses to the mathematics items in each of the four countries and found reasonably consistent support for a two-factor solution. For TIMSS2007, there was a slightly revised set of items and a new classification of items. Four items represented self-concept (the same four items as in the Index of "Students' self-confidence in learning" in TIMSS2003, sometimes referred to as self-confidence). Three items were used to represent a "Positive Affect" factor: The "enjoy" item from previous TIMSS studies (classified as "liking and competence" in one classification, but as "importance" in another classification for TIMSS2003) and two new items (liking and boring; see Appendix 1). One of the items (wanting to take more coursework) was not included in any of the three factors. Preuschoff, Martin and Mullis (personal communication, 20 October, 2011) noted that preliminary analyses indicated that this item was conceptually different and did not fit with any of the other factors.

In both 2007 and earlier data collections, TIMSS used what they called a "scale method" for multi-item scales for constructs that had an underlying quantitative continuum. Based on a simple average of responses of items (after reverse scoring negatively worded items) for each construct:

> TIMSS classified the students into three levels: high, medium, and low. In the International Reports, these derived variables are referred to as indices. To classify the cases into three groups, two cutoff points were established. Three main criteria were used in setting the cutoff points. First, the high level of the index should correspond to conditions or activities generally associated with good educational practice or high academic achievement. Second, there should be a reasonably even distribution of students across the three index levels. Third, the scale categories should be about the same size. (Ramirez & Arora, 2004, p. 315).

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

Comparisons of results across different countries and groups within countries require strong assumptions about the invariance of the factor structure across the groups or domains. If the underlying factors are fundamentally different, then there is no basis for interpreting observed differences (the "apples and oranges" problem). For example, in cross-national studies of self-concept and affect or even achievement differences, interpretations of mean differences—or even relations among different constructs—presuppose that the factors are the same across countries (i.e., Saudi and US students). In the present investigation, we consider a 2 (country) x 2 (gender) classification of invariance tests. More difficult are issues related to the invariance of item intercepts needed to justify the comparison of latent means. For example, assume that for four items designed to measure a particular trait, two clearly favor one country (or one gender) and two clearly favor the other country (or gender). These results provide no basis for evaluating mean differences in the trait, in that even the direction of differences would depend on the particular items used to measure the trait. Furthermore, because these 4 items are only a small sample of items that could be used to evaluate this trait, the results provide only a weak basis for knowing what would happen if a larger, more diverse sample of items were sampled. The invariance of item intercepts would mean that mean differences based on each of the items considered separately are reasonably consistent in terms of magnitude as well as direction, providing a much stronger basis of support for the generalizability of the interpretation of the observed mean differences.

Following Meredith (1993) and others (e.g., Marsh, Muthén, et al., 2009, Marsh, Lüdtke, Muthén et al., 2010) it is typical to consider a taxonomy of nested models that begins with a model with no invariance of any parameters, or *configural invariance*. The initial focus is on the invariance of the factor loadings—sometimes referred to as weak measurement invariance or pattern invariance—which requires that factor loadings be invariant over groups or over time. *Strong measurement invariance* requires that the indicator intercepts and factor loadings are invariant over groups (or domains), and justifies comparison of latent means. *Strict measurement invariance* requires the invariance of item uniquenesses (in addition to invariant factor loadings and intercepts), and justifies the comparison of manifest means over groups or time. Strict measurement invariance is required in order to compare manifest scale scores (or factor scores) in that differences in reliability for the multiple groups would distort mean differences on the observed scores. However, for comparisons based on latent constructs that are corrected for measurement error, the valid comparison of latent means only requires support for strong measurement invariance and not the additional assumption of the invariance of measurement error. Hence, comparison of group mean differences based on latent-variable models like those considered here makes fewer assumptions than that based on manifest scores. This is critical for analyses of TIMSS data in that there are substantial differences in the reliability of responses to the motivation constructs in different countries. Even less demanding is the comparison of latent correlations in different groups, a condition that only requires weak measurement invariance (i.e. invariance of the factor loadings). Although these tests require full invariance of all parameter estimates within each category (e.g., all factor loadings), Byrne, Shavelson and Muthén (1989) argued for the usefulness of a less demanding test of partial invariance in which a subset of parameter estimates are not constrained to be invariant.

Marsh, Abduljabbar et al. (2013) noted that there had apparently not been a rigorous evaluation of measurement invariance of TIMSS motivation items, which is a prerequisite for the comparison of means across countries or even groups within countries (e.g., gender differences) reported in the TIMSS documentation and the many cross-cultural secondary data analyses based on the TIMSS data. In response to these problems, Marsh, Abduljabbar et al. (2013) undertook a rigorous evaluation of the factor structure math and science motivation measures in four Arabic speaking countries (Saudi Arabia, Jordan, Oman, Egypt) and four English-speaking Anglo-Saxon countries (USA, England, Australia, Scotland). Consistent with previous research and a priori predictions, they found method effects associated with negatively worded items. After controlling for these method effects by including correlated uniquenesses between negatively worded items, they found good support for the invariance of factor loadings across all eight groups (i.e., weak measurement invariance). However, tests of the invariance of item intercepts across the eight countries were not completely satisfactory, although there was support for partial intercept invariance accomplished by allowing intercepts of some items to be freely estimated.

Single-Level Models of Measurement Invariance

In the present investigation and in cross-cultural research more generally, a rigorous evaluation of support for the a priori factor structure and the evaluation of measurement invariance is a critical prerequisite to the comparison of means and relations among variables across different countries. However, because these preliminary analyses are not the main focus of the present investigation, and they largely replicate the findings of Marsh, Abduljabbar, et al. (2013), they are presented here as Supplemental Materials that are readily available to readers. Nevertheless, the results and methodology are important to cross-cultural quantitative researchers. Following from Marsh, Abduljabbar, et al. (2013), we conducted preliminary analyses to evaluate measurement invariance across our four (2 gender x 2 country groups), with particular emphasis on the invariance of factor loadings (weak invariance), item intercepts (strong invariance) and also, invariance of item uniquenesses (strict invariance).

Traditionally, tests of invariance are based on two or more groups that vary in response to a single variable (analogous to a one-way analysis of variance). However, in the present investigation, our four groups vary along two dimensions (2 gender x 2 country; analogous to a two-way analysis of variance). Although it is straightforward to test invariance over all four groups, more nuanced comparisons are also of interest. Here, for example, it is also of interest to evaluate invariance over gender separately for each country without imposing invariance over country (gender within country) or, conversely, invariance across country separately for each gender (country within gender), or even invariance over all combinations of three groups, whilst not imposing invariance across the fourth group.

Our a priori model posits four factors: the two multi-item factors (self-concept and affect) and the two single-item factors (coursework aspiration and achievement; see Appendix). In our a priori model, we posited method effects based on responses to negatively worded items (see earlier discussion). In order to fully evaluate different aspects of this a priori model, we tested preliminary models of the negative item method effect and then subsequent models testing the invariance of the factor structure over responses by boys and girls from Saudi Arabia and US.

Non-invariant Factor Structures (Configural Invariance) and Method Effects. In the first set of models (M1a—M1e in Table 1) we explored the a priori factor structure with and without negative item method effects. M1 posited four factors, two based on multiple items (self-concept and affect) and two single-item constructs (achievement and aspirations). The fit of this model is reasonable (i.e., CFI =.956, but TLI= .931, RMSEA = .065). Consistent with a priori predictions, the inclusion of negative item method effects led to an improved fit (M1b). In M1c (Table 1), we required the correlated uniquenesses representing the negative item effect to be invariant over the four gender and country groups. The fit of Model 1C (CFI =.974, TLI= .958, RMSEA = .051) was very good and substantially improved, compared to Model M1a. Indeed, the improved fit of this model M1c, compared to the model M1a with no method effects, was substantial in relation to typical criteria of change in fit (ΔCFI =.018, ΔTLI= .027, ΔRMSEA = .014), but did not differ substantially from model M1b in which method effects were not constrained to be invariant over groups (ΔCFI =.004, ΔTLI= .002, ΔRMSEA = .002).

Weak measurement invariance (Factor loading invariance). In models considered thus far we did not impose any invariance constraints. For models in this section, we tested the invariance of factor loadings over four (2 gender x 2 countries) groups starting with Model M1c (with invariant correlated uniquenesses to control method effects). In Model M2a, we tested the invariance of factor loadings over all four groups and found a good fit to the data (CFI = .965, TLI = .952, RMSEA = .054). In subsequent models, we explored the effects of more restricted invariance constraints in relation to country (M2b) or to gender (M2c). However, neither of these submodels resulted in a substantial improvement in fit over the more parsimonious model M1a of complete factorial invariance over all four groups.

Factor loadings based on Model M2a are presented for all four countries in Table 2. Of course, in the unstandardized metric, the factor loadings for each indicator are all identical for all four groups. However, there are substantial differences in the sizes of the standardized factor loadings. Reflecting the negative item effect, the negatively worded items have systematically smaller factor loadings (see related findings for TIMSS2003 data in Chiu, 2011). However, particularly for math self-concept, standardized factor loadings are substantially lower for Saudi students than for US students (but similar across gender within each country). These results reflect the earlier observation that measurement error is substantially greater for Saudi students than for US students. In summary, the results provide reasonable support for the invariance of factor loadings across the four groups, but also reinforce why it is important to consider latent variable models that control for measurement error.

Strong Measurement Invariance (Item Intercept invariance). Strong measurement invariance requires that intercepts are invariant and is an important assumption that is needed to justify the comparison of latent means. In the most parsimonious models considered in this section, we tested the invariance of item intercepts over all four groups, but the fit was not very good (M3a in Table 1; CFI =.917, TLI= .899, RMSEA = .079). Separate tests of invariance across countries (M3c) or across genders (M3b) demonstrated that most of the misfit was due to a lack of invariance across countries. On the basis of modification indexes for Model M3a, we explored the possibility of partial intercept invariance of country, while assuming full intercept invariance over gender. Freeing invariance constraints across countries for the two negatively worded items from the self-concept factor resulted in a reasonable fit to the data (M3d in Table 1; CFI =.955, TLI= .943, RMSEA = .059). The fit of this model was better than M3a, with complete invariance of intercepts over all four groups (DCFI =.038, DTLI= .044, DRMSEA = .020). Because the metric of the latent means is established by the items with invariant intercepts, the results suggest that it is reasonable to compare latent mean differences between boys and girls within each country. However, because only 2 of 4 self-concept items had invariant intercepts over country, comparisons of latent mean self-concept scales across the two countries should be interpreted cautiously.

Strict Measurement Invariance (Item Uniqueness invariance). As noted earlier, reliability estimates are substantially higher for US responses than Saudi responses (see Appendix), so that invariance of measurement error across Saudi Arabia and the US does not exist for self-concept and positive affect scales. However, we also noted that strict measurement invariance is not a requirement for the comparison of latent means and other comparisons based on latent variable models such as those made in the present investigation. It is nevertheless relevant to evaluate this issue more formally within the context of multi-group tests of invariance over our four (2 gender x 2 countries) groups. Not surprisingly, adding the constraint of invariance of item uniquenesses across countries and genders to Model M3D as a test of strict measurement invariance resulted in a substantial decline in goodness of fit data (M4a in Table 1; CFI = .897, TLI = .949, RMSEA = .082). However, when we then evaluated whether this difference was a function of gender (invariance across gender; Model M4b) or country (invariance across countries), the results clearly demonstrated that the substantial differences in measurement were almost completely due to differences between the two countries rather than to gender differences. Although it might be possible to pursue partial invariance strategies, as we did with intercept invariance, we did not pursue this option because invariance of measurement error was not required for the latent variable models considered here. Nevertheless, these results demonstrate that comparisons between these two countries on self-concept and positive affect based on manifest scale scores are unwarranted, as are comparisons based on the trichotomized scores used to represent these factors in the TIMSS database.

Summary. In marked contrast to the technical sophistication in terms of achievement tests, there was little evidence of a strong theoretical basis for the selection and construction of self-concept and affect scales, consistent use of the same items over time, or even the presentation of psychometric evidence that has been incomplete, inadequate, or simply not presented (also see Liu & Meng, 2010). Consistent with previous TIMSS research, we found significant method effects associated with negatively worded items, so that CFAs that did not take these effects into account failed to provide an adequate fit to the data; these method effects are not easily incorporated into analyses based on manifest scores, which have been the basis of most secondary analyses with TIMSS and which can potentially invalidate cross-cultural comparisons. Once we controlled for a priori method effects, there was good support for invariance of factor loadings and intercepts over gender, and invariance of factor loadings over country. However, we found only partial invariance of item intercepts over country. Particularly when the number of items per factor is so small and complicated by negative-item effects, partial invariance provides a weak basis for making comparisons of latent means. Furthermore, cross-national comparisons of manifest means like those used in TIMSS reports, construction of the databases, and most secondary data analyses, require the further assumption that reliability and measurement error are invariant over country, but this assumption clearly was not met. From this perspective, the use of fully latent variables and preliminary analyses to valid the a prior factors structure, like those in used the present investigation, are critical for rigorous cross-cultural research.

Multi-Level Models of Measurement Invariance

Big-Fish-Little-Pond Effects of Class-average Achievement on Self-concept, Affect and Aspirations

The BFLPE predicts that individual student achievement has a positive effect on academic self-concept, but that class-average achievement has a negative effect on self-concept. The BFLPE is inherently a multilevel effect, juxtaposing the positive effects of achievement on self-concept at the individual student level (L1) and the negative effects of class-average achievement at the class (L2) level. In the present investigation, critical issues are the extent to which the BFLPE in relation to self-concept generalizes to positive affect and coursework aspirations, and these BFLPEs over country, gender, and the different constructs. Here we evaluate these issues with doubly-latent multilevel model. In this doubly-latent contextual model, the interpretation of contextual effects is facilitated by the invariance of factor loadings over the individual student (L1) and class-average (L2) levels. In preliminary analyses, we evaluated support for this cross-level invariance, extending the corresponding single-level Model M3d. The fit of Model M4a with factor loading invariance over level is good (see Table 1 in Supplemental Materials; CFI = .954, TLI = .941, RMSEA = .054) and did not differ substantially for less parsimonious models in which these cross-level invariance constraints were not imposed. To further facilitate interpretations, the BFLPE estimates that were in an unstandardized metric were then transformed into an effect size metric that was common across groups and the different constructs.

Supplemental Table 1

Summary of Goodness of Fit Statistics for Multigroup (MG) Single-level (SL) & Multilevel (ML) Models

Model	CHI	df	CFI	TLI	RMSEA	Description
Single-level: Configural Invariance and Negative Item Effect						
M1a	1240	92	.956	.931	.065	No negative item correlated uniqueness (NICU)s
M1b	659	80	.978	.960	.049	NICUs no invariance (Inv)
M1c	762	89	.974	.958	.051	NICUs Inv = Country (C) & Gender (G)
Single-level: Factor Loading Invariance (FL-Inv) based on Model M1c						
M2a	1009	104	.965	.952	.054	FL-Inv = Country (C) & Gender (G)
M2b	934	99	.968	.954	.053	FL-Inv = Country (Within each Gender)
M2c	919	99	.969	.954	.053	FL-Inv = Gender (within each country)
Eliminate factor-loading Invariance one group at a time based on Model M2a						
M2a1	905	99	.969	.955	.052	drop FL-Inv for Saudi Boys
M2a2	899	99	.969	.956	.052	drop FL-Inv for Saudi Girls
M2a3	974	99	.967	.951	.055	drop FL-Inv for US Boys
M2a4	966	99	.967	.952	.054	drop FL-Inv for US Girls
Single-level: Strong Invariance; Intercept Invariance (INT-Inv) based on Model M2a						
M3a	2294	119	.917	.899	.079	INT-INV = Country & Gender (full invariance)
M3b	1089	114	.963	.953	.054	INT-INV = Gender (Within each Country)
M3c	2272	114	.918	.896	.080	INT-INV = Country (Within each Gender)
M3d	1291	113	.955	.943	.059	INT-INV = Gender & Country (partial invariance)
Single-level: Strict Invariance; Uniqueness Invariance (UNQ-Inv) based on Model M2a						
M4a	2821	134	.897	.890	.082	UNQ-INV = Country & Gender (full invariance)
M4b	1295	127	.955	.949	.056	UNQ-INV = Gender (Within each Country)
M4c	2917	127	.893	.897	.086	UNQ-INV = Country (Within each Gender)
Multi-level: Invariance over student (Level 1) and classroom (Level 2) based on model M3D						
M5a	1994	225	.954	.941	.054	INV over levels (L1 = student, L2 = classroom) & four groups
M5b	1869	220	.957	.943	.050	INV over levels within groups
M5c	1644	205	.962	.947	.049	No multilevel invariance

Note. CHI = chi-square; df = degrees of freedom ratio; CFI = Comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation. CUs = a priori correlated uniquenesses based on the negatively worded items. All analyses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within schools. The results provide reasonable support for the invariance of factor loadings across the four (2 gender x 2 country) groups. Although the fit was slightly better in US samples, a model constraining the FLs to be equal across the 2 x 2 groups fitted well and was not substantially poorer than models allowing these to be freely estimated in each group. In a test of partial invariance of intercepts, invariance constraints were relaxed for two negatively-worded self-concept items, while constraining the intercepts to be invariant for the remaining self-concept items and all affect items. The results are clear, showing non-invariance of measurement error due to country rather than gender. Based on these preliminary analyses, Model 3D (full invariance of factor loadings, partial invariance of item intercepts) was the basis of subsequent analyses presented in the printed version of this article. For subsequent multilevel models we also demonstrated that factor loadings were reasonably invariant over multiple levels (students, Level 1; classrooms, Level 2).

Supplemental Table 2

Factor Structure for Saudi Arabia and United States: Factor Loadings

	Unstandardized	Standardized Solutions			
	All Groups	SA Girls	SA Boys	US Girls	US Boys
Variables	FLoad SE	FLoad SE	FLoad SE	FLoad SE	FLoad SE
Self-concept					
SCP1 [a]	1.000	0.482(.019)	0.468(.022)	0.810(.009)	0.816(.008)
SCN2	0.840(.019)	0.370(.016)	0.342(.015)	0.643(.012)	0.636(.011)
SCN3	1.019(.021)	0.480(.020)	0.449(020)	0.731(011)	0.727(.012)
SCP4	1.019(.010)	0.517(.022)	0.480(.023)	0.790(.008)	0.776(.009)
Affect					
AFP1[a]	1.000	0.831(.014)	0.819(.014)	0.878(.006)	0.869(.006)
AFN2	0.705(.014)	0.549(.013)	0.527(.014)	0.650(.012)	0.639(.013)
AFP3	0.995(.010)	0.829(.014)	0.812(.017)	0.916(.005)	0.901(.007)
Aspirations					
MORE [a]	1.000	1.000	1.000	1.000	1.000
Achievement					
ACH [a]	1.000	1.000	1.000	1.000	1.000

Note. See Appendix of the published article for a definition of the variables. These are average results over 5 data sets.

a Factor loadings were fixed to 1.0 to identify the model and thus have no standard errors for the unstandardized solution (or for the standardized solution for single-item factors).