Learning and Individual Differences 57 (2017) 141–156

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: Analysis of TIMSS 2011 data

Ze Wang*,[1], David A. Bergin[1]

University of Missouri, United States

ARTICLE INFO

Keywords:
Big-fish-little-pond effect
Cross-national research
Large-scale assessment
Multilevel latent variable modeling
TIMSS

ABSTRACT

This study used data from the Trends in International Mathematics and Science Study 2011 to examine the big-fish-little-pond effect (BFLPE) in eighth grade mathematics in 59 countries/regions. The BFLPE refers to the negative contrast effect of achievement on academic self-concept resulting from students comparing their ability to the average ability in their immediate context; students in higher-achieving contexts have lower average self-concept than similarly able students in lower-achieving contexts. Earlier research showed that controlling for students' perceived relative standing (how competent they perceive themselves compared to classmates) eliminates or reduces the BFLPE. However, little research has demonstrated the generalizability of this effect. Using the multilevel latent variable approach and large-scale assessment data, this study applied three statistical models. In the first model, measurement invariance of mathematics self-concept was assessed across countries/regions and across student and classroom levels within each country/region. In the second model, the BFLPE model was tested. The BFLPE was found in 58 countries/regions, with the Syrian Arab Republic being the only exception. In the third model, a perceived relative standing measure was added as a predictor at the within level to test whether perceived relative standing would eliminate or reduce the BFLPE. Results showed that the BFLPE was eliminated in six countries/regions and reduced in 49 countries/regions.

1. Introduction

This study examines the big-fish-little-pond effect (BFLPE) in 59 countries and regions using the Trends in International Mathematics and Science Study (TIMSS) 2011 database. The BFLPE refers to the pattern that students' academic self-concept tends to be depressed in settings where they perceive that their ability is lower than the ability of people immediately around them, even if their ability is high compared to the general population (Marsh, 1987). Depressed academic self-concept is important because academic self-concept is positively related to academic achievement, effort, aspirations, and school bonding (Ireson & Hallam, 2005; Jansen, Schroeders, Lüdtke, & Marsh, 2015; Marsh & Martin, 2011; Trautwein, Lüdtke, Schnyder, & Niggli, 2006b). Thus, influences that depress self-concept are important for educators to understand. The BFLPE is a robust pattern that has been demonstrated in many countries using the Program for International Student Assessment (PISA) and TIMSS data sets (e.g., Marsh et al., 2015; Nagengast & Marsh, 2012). The BFLPE depends upon social comparison (Marsh, Trautwein, Lüdtke, & Köller, 2008b; Marsh et al., 2008a); that is, students compare their own academic achievement to others' achievement in order to infer their own self-concept for the academic domain. However, in most of the cross-national research, social comparison is inferred, not measured. A purpose of the present research was to examine measured social comparison and the degree to which it amplified or reduced the BFLPE.

1.1. Self-concept

Self-concepts are people's perceptions of themselves as formed through their experiences with the environment and their interpretation of those experiences (Shavelson, Hubner, & Stanton, 1976). Students tend to differentiate their self-concepts in different academic subjects, for example, high self-concept for math but low for English (Jansen et al., 2015). Domain-specific self-concept and academic achievement in that domain correlate more strongly than self-concept for academic achievement in a different domain. For example, students' mathematics self-concept correlates strongly with mathematics grades and test scores, but global self-concept correlates weakly with mathematics performance (Swann, Chang-Schneider, & McClarty, 2007). In this study, we focus on mathematics self-concept.

* Corresponding author at: Department of Educational, School and Counseling Psychology, University of Missouri, 16 Hill Hall, Columbia, MO 65211, United States.
E-mail address: wangze@missouri.edu (Z. Wang).
[1] Department of Educational, School and Counseling Psychology, University of Missouri.

http://dx.doi.org/10.1016/j.lindif.2017.04.003
Received 8 October 2015; Received in revised form 3 February 2017; Accepted 3 April 2017

1.2. Big-fish-little-pond effect

The BFLPE was proposed to help explain the relationship between self-concept in an academic domain and academic achievement in that domain (Marsh, 1987; Marsh & Parker, 1984). Academic self-concepts are influenced by personal achievement and by the level of surrounding achievement (e.g., Liem, Marsh, Martin, McInerney, & Yeung, 2013). One could say that there are two distinct influences on students' self-concept. First, students' own accomplishments and perceptions of their own ability influence their self-concept. If individual students' experience suggests that they are good at particular activities, students tend to have high self-concept for those activities (Guay, Marsh, & Boivin, 2003). Second, students' perceptions of the range of competence around them influence their self-concept. Thus, the second influence on self-concept is social comparison of competence relative to others (Marsh et al., 2008b).

In earlier research, the BFLPE was referred to as the negative effect of the school (or classroom) average achievement on self-concept after controlling for the effect of individual achievement on self-concept (e.g., Marsh, 1987). However, this earlier operationalization did not explicitly model the relationship between the school/classroom average achievement and the school/classroom average self-concept. In recent research using multilevel modeling (e.g., Arens & Watermann, 2015; Marsh et al., 2015), the BFLPE is represented by a negative difference between a positive level 2 (e.g., school or classroom level) effect (between-level effect) and a positive level 1 (i.e., student level) effect (within-level effect). In a relatively constrained environment such as the classroom, a higher-achieving student would have higher self-concept, compared to other students in the same context; thus a positive relationship between achievement and self-concept exists at the student level (level 1 or within level). A positive achievement-self-concept relationship is also expected to exist at the between school/classroom level (level 2 or between level) (Wang, 2015). It is assumed that the positive effect at the between level would be smaller than the within-level effect and their difference, which is expected to be negative, is the BFLPE (Marsh et al., 2012; Marsh et al., 2015).

1.3. Perceived relative standing

Dai and Rinn (2008) critiqued BFLPE research, pointing out that the BFLPE is based on the assumption of social comparison, but that social comparison is typically not explicitly measured by asking students how they compare to fellow students. Marsh and colleagues acknowledged this (Marsh et al., 2008a). Studies that use international data sets (e.g., PISA and TIMSS) have used implied social comparison; that is, they ask students to rate their self-concept and then the researchers compare students' self-concepts across classrooms or schools that have different average achievement while controlling for individual student achievement. They then infer that the students are engaged in social comparison because students in high-achieving schools/classrooms report lower self-concept compared to similarly achieving students in lower-achieving schools/classrooms. The data structure of TIMSS allows comparisons in classrooms and PISA in schools.

A study that explicitly measured students' perceived relative standing (how competent they perceived themselves compared to classmates) eliminated the BFLPE (Huguet et al., 2009). The authors tested the assumption that social comparison with the entire class underlies the BFLPE and therefore the BFLPE should be eliminated when students' perceived relative standing in the class is controlled. They had grade 6 French students rate their competence compared to their classmates. The analysis rendered the BFLPE nonsignificant for French and math classes when perceived relative standing was included. The BFLPE effect changed from -0.47 ($p < 0.001$) to -0.07 (ns) for math and -0.45 ($p < 0.001$) to -0.05 (ns) for French. These findings supported their assumption. In another study, using TIMSS 2007, Wang (2015) found that out of the 45 countries that had a statistically significant BFLPE ($p < 0.01$), 43 countries had a smaller BFLPE after adding a perceived relative standing measure, but the BFLPE was not eliminated in any of the countries. Because Wang analyzed far more data than Huguet et al., the consistent finding seems to be that the BFLPE is reduced but not eliminated by controlling for perceived relative standing. The present study was a further test of this pattern of results.

1.4. The present study

In the present study, we explicitly measured social comparison through an item that addressed perceived relative standing. This provides a test of whether the BFLPE is predicated upon student comparisons with classmates. We followed Wang (2015) by using an item that asked students whether mathematics was difficult for them compared to their classmates.

This study is a replication and extension to new data and new countries/regions of Wang's research. Wang (2015) used the TIMSS 2007 database to study the BFLPE and explicit social comparison in 49 countries. The present study used the TIMSS 2011 database, which represents 59 countries/regions at eighth grade; 22 countries/regions were new compared to Wang (2015). Thus the present study utilized different students in a different year in 22 additional sites. Researchers in psychology have placed increasing importance on replication as evidenced by the 2012 special issue on replicability in Perspectives on Psychological Science (see also Lindsay, 2015). There is increasing effort to replicate published studies in order to understand the degree to which psychological phenomena are consistent; for example, researchers who replicated 100 published experimental and correlational studies found that while 97% of the original studies had statistically significant results, only 36% of replications had statistically significant results (Open Science Collaboration, 2015). Average effect sizes (r) were reduced from 0.403 to 0.197. This finding highlights the importance of verifying whether research results are reliable across time and across samples.

2. Methods

2.1. Sample and variables

The present study used TIMSS 2011 data from 59 countries and regions at the eighth-grade level (including three countries whose students were not in eighth grade: Botswana - 6th and 9th, Honduras - 6th and 9th, and South Africa - 9th) (Foy, Arora, & Stanco, 2013). The total sample consisted of 306,048 students from 12,519 classes and 9739 schools (see Table 1). Mathematics self-concept was measured by three items: (a) I usually do well in mathematics; (b) mathematics is not one of my strengths; and (c) I learn things quickly in mathematics. Perceived relative standing was measured by one item: mathematics is more difficult for me than for many of my classmates. While Marsh et al. (2014) and Marsh et al. (2015) used this item as part of their measure of self-concept, it has clear social comparison aspects that make it appropriate for examining perceived relative standing. One can see that it is quite different from the other three items, which do not ask for a social comparison. We also used factor analysis to show that the perceived relative standing item was different from the other three items. See Appendix A.

The three mathematics self-concept items and the one perceived relative standing item were rated on a 1 to 4 Likert-scale, and reverse coded when appropriate. A higher value on the items means a higher level of mathematics self-concept or more positively perceived relative standing. For mathematics achievement, TIMSS 2011 used a matrix sampling technique, and students answered questions in item blocks. Students' mathematics achievement was estimated using item response theory together with a multiple imputation technique. Each student's mathematics achievement was measured by five plausible values. After

Table 1
Results for Model 1 - multilevel confirmatory factor analysis with multilevel measurement invariance of mathematics self-concept in 59 countries/regions.

Country/Region	# Schools	# Classes	# Students	Chi-square	CFI	TLI	RMSEA	Variance of mathematics self-concept (within-class)	Variance of mathematics self-concept (between-class)	ICC
Abu Dhabi, UAE	166	195	4373	2.17	1.000	0.999	0.004	0.284***	0.039***	0.121
Dubai, UAE	130	255	5571	8.69*	0.995	0.986	0.025	0.493***	0.069***	0.123
United Arab Emirates	458	628	14,089	9.11*	0.996	0.989	0.016	0.343***	0.050***	0.127
Armenia	153	251	5846	13.97***	0.994	0.981	0.032	0.745***	0.042***	0.053
Australia	275	332	6566	29.29***	0.993	0.978	0.046	0.584***	0.117***	0.167
Bahrain	95	181	4640	0.16	1.000	1.008	0.000	0.326***	0.022**	0.063
Botswana (6th and 9th)	150	151	5400	4.60	0.998	0.993	0.016	0.439***	0.048***	0.099
Alberta, Canada	145	221	4799	0.42	1.000	1.001	0.000	0.604***	0.037***	0.058
Chile	193	194	5835	1.80	1.000	1.000	0.000	0.631***	0.031***	0.047
Ontario, Canada	143	243	4756	1.76	1.000	1.000	0.000	0.690***	0.026**	0.036
Quebec, Canada	189	263	6149	10.35**	0.998	0.995	0.026	0.615***	0.065***	0.096
England	118	176	3842	54.87***	0.972	0.915	0.083	0.464***	0.161***	0.258
Finland	145	258	4266	3.19	1.000	0.999	0.012	0.740***	0.050***	0.063
Georgia	172	200	4563	5.73	0.997	0.992	0.020	0.663***	0.043***	0.061
Ghana	161	173	7323	1.18	1.000	1.003	0.000	0.457***	0.053***	0.104
Hong Kong SAR	117	121	4015	22.15***	0.989	0.968	0.050	0.631***	0.032**	0.048
Honduras (6th and 9th)	155	160	4418	1.68	1.000	1.001	0.000	0.541***	0.075***	0.122
Hungary	146	251	5178	3.49	1.000	0.999	0.012	0.639***	0.043***	0.063
Indonesia	153	174	5795	4.99	0.996	0.989	0.016	0.363***	0.115***	0.241
Iran, Islamic Rep. of	238	239	6029	0.21	1.000	1.003	0.000	0.476***	0.058***	0.109
Israel	151	169	4699	1.20	1.000	1.001	0.000	0.595***	0.040***	0.063
Italy	197	205	3979	0.27	1.000	1.002	0.000	0.695***	0.019**	0.027
Jordan	230	253	7694	2.32	0.999	0.998	0.005	0.164***	0.014**	0.079
Japan	138	139	4414	0.86	1.000	1.001	0.000	0.676***	0.013*	0.019
Kazakhstan	147	206	4390	0.29	1.000	1.004	0.000	0.481***	0.065***	0.119
Korea, Rep. of	150	150	5166	6.91*	0.998	0.995	0.022	0.806***	0.015**	0.018
Lebanon	147	187	3974	0.87	1.000	1.004	0.000	0.532***	0.054***	0.092
Lithuania	141	258	4747	16.33***	0.994	0.983	0.039	0.726***	0.055***	0.070
Morocco	279	279	8986	7.12*	0.995	0.984	0.017	0.143***	0.014***	0.089
Macedonia, Rep. of	150	204	4062	1.34	1.000	1.001	0.000	0.558***	0.063***	0.101
Malaysia	180	180	5733	1.59	1.000	1.001	0.000	0.470***	0.078***	0.142
Norway	134	170	3862	10.26**	0.998	0.993	0.033	0.665***	0.022*	0.032
New Zealand	158	241	5336	10.42**	0.996	0.989	0.028	0.608***	0.071***	0.105
Oman	323	364	9542	1.85	1.000	1.001	0.000	0.026	0.003	0.103
Palestinian Nat'l Auth.	201	247	7812	1.63	1.000	1.001	0.000	0.318***	0.050***	0.136
Qatar	109	195	4422	0.30	1.000	1.009	0.000	0.244***	0.037**	0.132
Romania	147	248	5523	0.83	1.000	1.002	0.000	0.668***	0.058***	0.080
Russian Federation	210	229	4893	7.83*	0.998	0.994	0.024	0.579***	0.047***	0.075
Saudi Arabia	153	164	4344	3.85	0.998	0.994	0.015	0.297***	0.046***	0.134
Singapore	165	330	5927	3.03	1.000	0.999	0.009	0.672***	0.078***	0.104
Slovenia	186	225	4415	1.68	1.000	1.000	0.000	0.666***	0.028***	0.040
Sweden	153	266	5573	0.50	1.000	1.001	0.000	0.624***	0.014**	0.022
Syrian Arab Republic	148	148	4413	0.60	1.000	1.010	0.000	0.109***	0.013***	0.107
Thailand	172	172	6124	3.14	0.999	0.997	0.010	0.457***	0.051***	0.100
Tunisia	207	207	5128	9.42**	0.990	0.970	0.027	0.171***	0.017***	0.090
Turkey	239	240	6928	6.67*	0.999	0.997	0.018	0.652***	0.058***	0.082
Chinese Taipei	150	152	5042	1.18	1.000	1.001	0.000	0.708***	0.046***	0.061
Alabama, US	55	107	2113	139.53***	0.930	0.789	0.182	0.536***	0.073***	0.120
California, US	82	110	2614	11.17**	0.993	0.979	0.042	0.543***	0.059***	0.098
Colorado, US	53	105	2167	25.40***	0.977	0.932	0.074	0.539***	0.097***	0.153
Connecticut, US	62	120	2099	107.78***	0.935	0.805	0.161	0.511***	0.091***	0.151
Florida, US	60	113	1712	18.31***	0.985	0.955	0.070	0.541***	0.094***	0.148
Indiana, US	56	114	2260	30.13***	0.982	0.947	0.079	0.530***	0.106***	0.167
Ukraine	148	162	3378	34.22***	0.971	0.912	0.069	0.642***	0.095***	0.129
Massachusetts, US	56	105	2075	32.58***	0.985	0.954	0.086	0.560***	0.084***	0.130
Minnesota, US	55	110	2500	43.84***	0.983	0.948	0.092	0.563***	0.110***	0.163
North Carolina, US	59	105	2103	6.39*	0.996	0.989	0.032	0.513***	0.097***	0.159
United States	501	557	10,477	83.99***	0.989	0.967	0.063	0.519***	0.092***	0.151
South Africa (9th)	285	317	11,969	7.89*	0.994	0.983	0.016	0.392***	0.056***	0.125

Countries are arranged in alphabetic order of country code, but country code is not shown in the table. There are three out-of-grade countries whose students were not in eighth grade: Botswana, Honduras, and South Africa. Multilevel measurement invariance in the multilevel confirmatory factor analysis model refers to factor loading invariance across the student and classroom levels. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; ICC = intraclass correlation coefficient.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

the data management step, the five plausible values were treated as imputed values (L.K. Muthén & Muthén, 2012). It is very important to consider the complex structure of data collected in TIMSS. Appendix A describes how we did this. Appendix B includes a SAS macro we wrote for data management and codes for obtaining descriptive statistics.

2.2. Statistical models

The BFLPE in recent studies using multilevel latent variable modeling is operationalized as the difference between the between-level effect and the within-level effect with the predictor variable group mean centered. The group mean can be either the manifest aggregation of student achievement (e.g., Marsh et al., 2015) or the latent aggregation of student achievement (e.g., Marsh et al., 2012; Wang, 2015). The latent aggregation is the between-level component of the decomposition of variables in Eqs. (1)–(4) presented below. The latent aggregation results in implicit, instead of observed, group-mean centering and unbiased estimates of contextual effects (Lüdtke et al., 2008), although the increased model complexity associated with the latent aggregation may lead to less stable parameter estimates (Marsh et al., 2009).

In order to separate the contextual effect from the internal, person effect, both the predictor variable (i.e., mathematics achievement in this study) and the outcome variable (i.e., mathematics self-concept in this study) can be thought of as consisting of two components. The first component, the within-class component, is differences among students within the same class. The within-class component of mathematics achievement reflects the forced relative standing of a student in the classroom since all students in the same classroom are compared to the same classroom average. The within-class component of mathematics achievement affects the within-class component of mathematics self-concept. The second component, the between-class component, represents differences between classes. The decomposition of each variable into a within and a between component, conceptually, is the same as in random-effect analyses of variance (ANOVAs), except that the mathematics self-concept variable is latent. For mathematics self-concept, the decomposition is:

$$\mathrm{sc_{ij}} = \mathrm{sc_{wij}} + \mathrm{sc_{bj}} \tag{1}$$

where $\mathrm{sc_{ij}}$ is the latent mathematics self-concept for student *i* in class *j*; $\mathrm{sc_{wij}}$ is its within component and $\mathrm{sc_{bj}}$ is its between component. The grand mean of $\mathrm{sc_{ij}}$ is arbitrary and set to zero. $\mathrm{sc_{wij}}$ and $\mathrm{sc_{bj}}$ are measured by three indicators with the following equations:

$$\mathbf{Y_{ij}} = \boldsymbol{\alpha}_\mathbf{j} + \boldsymbol{\lambda}\mathrm{sc_{wij}} + \mathbf{e_{ij}} \tag{2}$$

$$\boldsymbol{\alpha}_\mathbf{j} = \boldsymbol{\gamma} + \boldsymbol{\lambda}\mathrm{sc_{bj}} + \mathbf{r_j} \tag{3}$$

For each student, $\mathbf{Y_{ij}}$ is a vector with three elements, representing values of the three mathematics self-concept items. $\boldsymbol{\alpha}_\mathbf{j}$ is a vector representing class-specific indicator intercepts at the within level that serve as indicators of the latent factor $\mathrm{sc_{bj}}$ at the between level; γ is a vector for the grand mean indicator intercepts that are the same for all groups (i.e., classes in this study). The loadings, in the vector $\boldsymbol{\lambda}$, corresponding to each observed variable Y, are invariant across levels. This ensures that the interpretation of mathematics self-concept at the within and between levels are similar.

2.2.1. Model 1

The first statistical model is a two-level confirmatory factor analysis (CFA) model for mathematics self-concept (see Fig. 1). The solid dots at the end of the directional arrows from $\mathrm{sc_{wij}}$ to its indicators suggest random intercepts for different classes. These random intercepts, α_{1j}, α_{2j}, and α_{3j}, are modeled at the between-level.

2.2.2. Model 2

Similar to how mathematics self-concept is decomposed in Eq. (1), student mathematics achievement is decomposed as below:

$$\mathrm{math_{ij}} = \mu_{math} + \mathrm{math_{wij}} + \mathrm{math_{bj}} \tag{4}$$

where $\mathrm{math_{ij}}$ is the mathematics achievement of student *i* in class *j*; μ_{math} is the grand mean of mathematics achievement for all students in all classes and is a constant; $\mathrm{math_{wij}}$ is student *i*'s mathematics achievement around the class-average mathematics achievement; and $\mathrm{math_{bj}}$ is the average mathematics achievement for class *j*, around the grand mean.

Next, it is assumed that the within and between parts are orthogonal. That is, the within-class component of mathematics achievement only affects the within-class component of mathematics self-concept; and the between-class component of mathematics achievement only affects the between-class component of mathematics self-concept. The between-class component of mathematics achievement is modeled as a latent variable to account for the measurement errors associated with observed class averages of mathematics achievement (L. K. Muthén & Muthén, 2012), which leads to the latent aggregation of student achievement to form class average achievement. The statistical model is below:

$$\mathrm{sc_{wij}} = \beta_{within}\mathrm{math_{wij}} + \beta_2(\mathrm{math_{ij}})^2 + \varepsilon_{ij} \tag{5}$$

$$\mathrm{sc_{bj}} = \beta_{between}\mathrm{math_{bj}} + \delta_j \tag{6}$$

This is an extension of the multilevel latent covariate model proposed by Lüdtke et al. (2008) by using a latent factor as the outcome variable. The difference between $\beta_{between}$ and β_{within} is the BFLPE. β_2 is the coefficient for the quadratic component of student mathematics achievement, which is often included in BFLPE research in addition to the linear component (e.g., Marsh & Hau, 2003). The figurative representation is in Fig. 2. Assuming that ε_{ij} is unrelated to $\mathrm{math_{wij}}$ or $(\mathrm{math_{ij}})^2$ (i.e., their covariances are zero), and that $\mathrm{math_{bj}}$ and δ_j are unrelated, it can be derived that the variances of $\mathrm{sc_{wij}}$ and of $\mathrm{sc_{bj}}$, respectively, are

$$\begin{aligned}\mathrm{Var}(\mathrm{sc_{wij}}) = {} & \beta_{within}^2 \times \mathrm{Var}(\mathrm{math_{wij}}) + \beta_2^2 \times \mathrm{Var}((\mathrm{math_{ij}})^2) + 2\beta_{within}\beta_2\mathrm{Cov} \\ & (\mathrm{math_{wij}}, \mathrm{math_{ij}^2}) + \mathrm{Var}(\varepsilon_{ij})\end{aligned} \tag{7}$$

$$\mathrm{Var}(\mathrm{sc_{bj}}) = \beta_{between}^2 \times \mathrm{Var}(\mathrm{math_{bj}}) + \mathrm{Var}(\delta_j) \tag{8}$$

The variance of $\mathrm{sc_{ij}}$ is therefore,

$$\begin{aligned}&\mathrm{Var}(\mathrm{sc_{ij}}) = \mathrm{Var}(\mathrm{sc_{wij}}) + \mathrm{Var}(\mathrm{sc_{bj}}) \\ &= \beta_{within}^2 \times \mathrm{Var}(\mathrm{math_{wij}} + \beta_2^2 \times \mathrm{Var}((\mathrm{math_{ij}})^2) \\ &\quad + 2\beta_{within}\beta_2\mathrm{Cov}(\mathrm{math_{wij}}, \mathrm{math_{ij}^2}) + \mathrm{Var}(\varepsilon_{ij}) \\ &\quad + \beta_{between}^2 \times \mathrm{Var}(\mathrm{math_{bj}}) + \mathrm{Var}(\delta_j\end{aligned} \tag{9}$$

Standardized effect sizes were calculated following recent BFLPE studies (Marsh et al., 2009; Marsh et al., 2015; Nagengast & Marsh, 2012). At the within- and between- levels, effect sizes can be calculated respectively as:

$$\mathrm{ES_{within}} = 2 \times \beta_{within} \times \mathrm{SD}(\mathrm{math_{wij}})/\mathrm{SD}(\mathrm{sc_{ij}}) \tag{10}$$

$$\mathrm{ES_{between}} = 2 \times \beta_{between} \times \mathrm{SD}(\mathrm{math_{bj}})/\mathrm{SD}(\mathrm{sc_{ij}}) \tag{11}$$

where $\mathrm{SD}(\mathrm{math_{wij}})$and $\mathrm{SD}(\mathrm{math_{bj}})$ are the standard deviations of the within-level and between-level mathematics achievement, respectively, and $\mathrm{SD}(\mathrm{sc_{ij}})$ is the standard deviation of the total mathematics self-concept calculated as

$$\mathrm{SD}(\mathrm{sc_{ij}}) = \sqrt{\mathrm{Var}(\mathrm{sc_{ij}})} \tag{12}$$

Since the BFLPE is a contextual effect, its effect size is in relation to the between-level variance of mathematics achievement and is therefore calculated as:

$$\mathrm{ES_{BFLPE}} = 2 \times (\beta_{between} - \beta_{within}) \times \mathrm{SD}(\mathrm{math_{bj}})/\mathrm{SD}(\mathrm{sc_{ij}}) \tag{13}$$

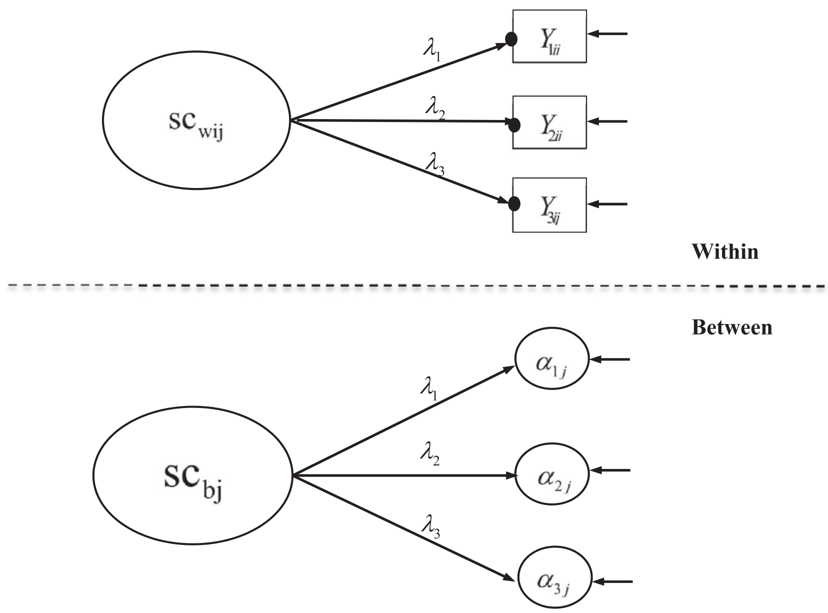

Fig. 1. Statistical Model 1 - two-level confirmatory factor analysis model with multilevel measurement invariance of mathematics self-concept.

2.2.3. Model 3

In the third statistical model, perceived relative standing is added as an additional predictor of the within-level mathematics self-concept, consistent with previous studies (Huguet et al., 2009; Wang, 2015). Although it is possible to model the effect of perceived relative standing at both levels, theoretically, it makes sense to include it only at the within-level. Perceived relative standing was a within-level measure because it asked explicitly that the students compare themselves to their classmates. Fig. 3 illustrates this model.

For Model 1, a total sample analysis and a multi-group analysis were conducted with data from all 59 countries/regions together before separate analyses were done for individual countries/regions. Based on results from Model 1, for Models 2 and 3 separate analyses were conducted for individual countries/regions. For individual country/region analyses, three Mplus templates were written, one for each statistical model mentioned earlier. These Mplus templates are in Appendix C. The R package "MplusAutomation" (Hallquist, 2011) was used to create and run Mplus syntax files for individual countries. Parameter estimates for individual countries/regions were extracted from Mplus output files. The R codes for using the package "MplusAutomation" to create, run, and extract results from Mplus models are in Appendix D.

3. Results

Table 1 presents results for Model 1 for separate countries/regions. Table 2 presents results for Models 2 and 3 for separate countries/regions.

3.1. Multilevel CFA of mathematics self-concept (Model 1)

We conducted a total sample analysis and a multi-group analysis using all data from the 59 countries/regions together. Based on the regular model fit cutoffs (root mean square error of approximation, or RMSEA < 0.08; Browne & Cudeck, 1993; comparative fit index, or CFI > 0.95, Tucker–Lewis index, or TLI > 0.95; Hu & Bentler, 1998, 1999), the model fit for the total sample analysis was good: CFI = 0.999; TLI = 0.997, RMSEA = 0.005. Using country/region as the grouping variable, for the multi-group analysis, the configural invariance model fit the data perfectly because it was a saturated model. However, the metric invariance model (i.e., the model with factor loading invariance across countries/regions) did not fit the data very well based on RMSEA and TLI: RMSEA = 0.095; CFI = 0.950; TLI = 0.924; and the metric invariance model fit worse than the configural invariance model (ΔRMSEA = 0.095; ΔCFI = 0.050; ΔTLI = 0.076) (Chen, 2007; Cheung & Rensvold, 2002). These results

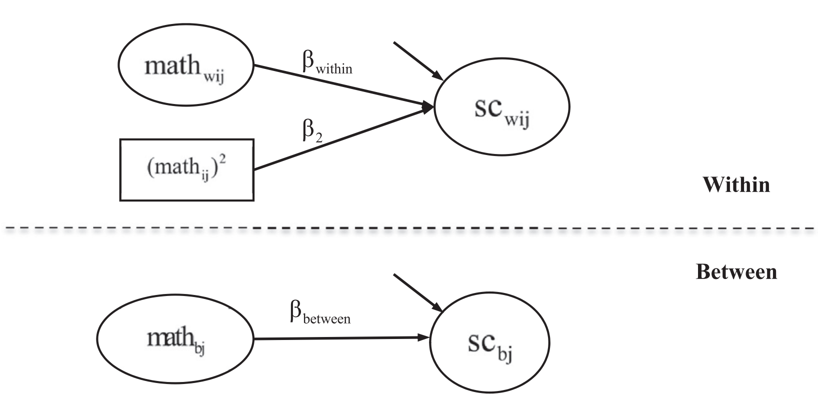

Fig. 2. Statistical Model 2 - model to test the big-fish-little-pond effect. Indicators of the within- and between-level mathematics self-concept are not shown in figure.

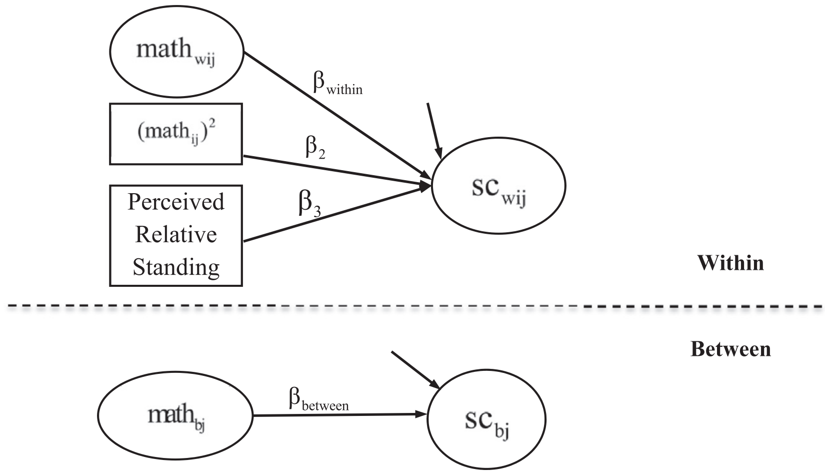

Fig. 3. Statistical Model 3 - model to test the big-fish-little-pond effect after controlling for the perceived relative standing. Indicators of the within- and between-level mathematics self-concept are not shown in figure.

mean that the mathematics self-concept factor and the indicators did not have equivalent relationships across countries/regions. Metric invariance is a precondition for testing invariance of relations between constructs. Because the metric invariance model did not fit the data very well, when analyzing relationships between mathematics self-concept and other variables, analyses should be conducted separately for each country/region.

We tested Model 1 for each of the 59 countries/regions separately by applying the multilevel (i.e., two-level with the COMPLEX TWOLEVEL type of analysis in Mplus) CFA model in each country/region. The model fit indices of Model 1 for the 59 countries/regions are in Table 1. The model did not fit data from six US states (Alabama, Colorado, Connecticut, Indiana, Massachusetts, and Minnesota) and two countries (England and Ukraine). However, only two of these eight cases (Alabama and Connecticut) had very poor model fit (for Alabama, CFI = 0.930, TLI = 0.789, and RMSEA = 0.182; for Connecticut, CFI = 0.935, TLI = 0.805, and RMSEA = 0.161). The other six cases had CFI > 0.95, TLI > 0.91, and RMSEA close to or < 0.08. Table 1 also includes the within-level variance, the between-level variance, and the intraclass correlation coefficient (ICC) of mathematics self-concept for each country/region. The smallest ICC was in Korea (1.8%), followed by Japan (1.9%) and the largest was in England (25.8%), followed by Indonesia (24.1%). A small ICC means that there is little between class variation compared to within class. However, a small ICC of mathematics self-concept can also be viewed as resulting from social comparison largely within the class.

3.2. BFLPE (Model 2)

The effect size of the BFLPE was specified as an additional parameter, and hypothesis testing of this parameter was a test for the presence of the BFLPE.

The within-level effect of mathematics achievement on mathematics self-concept,β_{within}, was positive and statistically significant ($p < 0.001$ in all 59 countries/regions). At the 0.05 level, the between-level effect of mathematics achievement on mathematics self-concept, $\beta_{between}$, was not statistically significant in 12 countries/regions. The effect was positive and statistically significant in 43 countries/regions. The mean effect size in these 43 countries/regions was 0.35, measured with Cohen's *d* and the median was 0.28. The range was from 0.12 (Sweden) to 0.91 (Minnesota - US). The effect was negative and statistically significant in four countries (Botswana, Honduras, Indonesia and South Africa). The difference between $\beta_{between}$ and β_{within} (i.e., the BFLPE) was negative and statistically significant in all but one country (the Syrian Arab Republic). The mean and median effect sizes were − 0.59. Aside from Syria, the range of effect sizes was from − 0.24 (in Slovenia, a small effect) to − 1.33 (in South Africa, a very large effect). We did not note patterns in magnitudes of the BFLPE based on geographical setting or ethnic/religious group. For example, predominantly Muslim countries ranged in effect size from − 0.11 (not statistically significant) for Syria to − 0.74 for United Arab Emirates. European countries ranged from − 0.24 for Slovenia to − 0.78 for Hungary. East and Southeast Asian countries ranged from − 0.26 for Chinese Taipei to − 1.25 for Singapore. Even states within the United States of American ranged from − 0.35 for Indiana to − 0.81 for Alabama. The effect sizes of β_{within}, $\beta_{between}$, and the BFLPE are shown in Table 2.

3.3. Perceived relative standing and BFLPE on mathematics self-concept (Model 3)

The third model tests whether the BFLPE was eliminated or reduced after controlling for students' perceived relative standing among classmates.

The effect of perceived relative standing was positive and statistically significant in all countries/regions. Effect sizes are in Table 2. The average effect size was 0.91 with a median effect size of 1.01. The range was 0.23 (Honduras) to 1.31 (Chinese Taipei).

Controlling for perceived relative standing, the within-level effect of mathematics achievement on mathematics self-concept, β_{within}, was positive and statistically significant in all countries ($p < 0.001$). At the 0.05 level, the between-level effect of mathematics achievement on mathematics self-concept, $\beta_{between}$, was not statistically significant in 18 countries/regions, was positive and statistically significant in 35 countries/regions, and was negative and statistically significant in six countries; the BFLPE controlling for perceived relative standing was statistically significant in 52 countries/regions. Of the seven countries/regions where there was no BFLPE after controlling for perceived relative standing, six had a statistically significant BFLPE before controlling for perceived relative standing. In other words, in these six countries/regions (England, Chinese Taipei, Florida-US, Indiana-US, Minnesota-US, and North Carolina-US), including perceived relative standing eliminated the BFLPE. In 49 countries/regions, the BFLPE was smaller after controlling for perceived relative standing.

Interestingly, after controlling for perceived relative standing, the effect of mathematics achievement at the between level decreased in all but one country (Jordan). In Jordan, the effect changed from 0.20 in Model 2 to 0.24 in Model 3. In the four countries (Botswana, Honduras, Indonesia, and South Africa) where a negative effect of mathematics achievement at the between level existed in Model 2, these negative effects still existed and became stronger in Model 3. In two countries

Table 2
Effect sizes in 59 countries/regions.

Country/Region	Model 2			Model 3			
	Within-level linear effect of mathematics achievement on MSC	Between-level effect of mathematics achievement on MSC	Contextual effect (BFLPE)	Within-level linear effect of mathematics achievement on MSC after controlling for PRS	Between-level effect of mathematics achievement on MSC after controlling for PRS)	Effect of PRS	Contextual effect (BFLPE) after controlling for PRS
Abu Dhabi, UAE	0.90***	0.09	−0.60***	0.63***	0.08	1.14***	−0.35***
Dubai, UAE	0.96***	0.03	−0.95***	0.64***	−0.14*	0.86***	−0.75***
United Arab Emirates	0.90***	0.02	−0.74***	0.64***	−0.08	0.99***	−0.57***
Armenia	0.89***	0.19**	−0.33***	0.73***	0.15*	0.55***	−0.26***
Australia	1.10***	0.66***	−0.76***	0.63***	0.41***	1.08***	−0.34***
Bahrain	0.97***	0.15	−0.65***	0.71***	−0.02	1.14***	−0.55***
Botswana (6th and 9th)	0.51***	−0.24***	−0.49***	0.37***	−0.30***	0.46***	−0.48***
Alberta, Canada	1.32***	0.28***	−0.47***	0.69***	0.12**	1.18***	−0.25***
Chile	1.20***	0.15***	−0.91***	1.01***	0.02	0.48***	−0.81***
Ontario, Canada	1.34***	0.28***	−0.51***	0.68***	0.11**	1.20***	−0.25***
Quebec, Canada	1.22***	0.38***	−0.70***	0.63***	0.16***	1.20***	−0.36***
England	0.68***	0.77***	−0.47***	0.36***	0.62***	1.04***	−0.01
Finland	1.32***	0.34***	−0.47***	0.77***	0.19***	1.03***	−0.23***
Georgia	1.12***	0.26***	−0.50***	0.94***	0.19**	0.50***	−0.41***
Ghana	0.55***	−0.05	−0.51***	0.46***	−0.15	0.58***	−0.51***
Hong Kong SAR	1.06***	0.36***	−0.93***	0.60***	0.12*	1.19***	−0.58***
Honduras (6th and 9th)	0.62***	−0.21*	−0.68***	0.56***	−0.24**	0.23***	−0.64***
Hungary	1.33***	0.17*	−0.78***	0.91***	0.03	0.85***	−0.57***
Indonesia	0.33***	−0.47***	−0.73***	0.15***	−0.52***	1.01***	−0.65***
Iran, Islamic Rep. of	0.95***	0.25***	−0.61***	0.74***	0.14*	0.65***	−0.50***
Israel	1.00***	0.23***	−0.58***	0.60***	0.02	1.07***	−0.41***
Italy	1.25***	0.03	−0.66***	0.82***	0.02	0.92***	−0.43***
Jordan	0.92***	0.20**	−0.37***	0.87***	0.24**	1.02***	−0.27***
Japan	1.30***	0.12*	−0.45***	0.95***	0.09	0.92***	−0.32***
Kazakhstan	0.70***	0.09	−0.61***	0.42***	−0.02	1.21***	−0.43***
Korea, Rep. of	1.41***	0.18***	−0.27***	0.90***	0.15***	0.96***	−0.12***
Lebanon	0.86***	0.18**	−0.56***	0.70***	0.07	0.61***	−0.52***
Lithuania	1.32***	0.13*	−0.63***	0.89***	0.05	0.89***	−0.45***
Morocco	0.64***	0.15**	−0.29***	0.59***	0.11*	1.08***	−0.29***
Macedonia, Rep. of	0.85***	−0.04	−0.75***	0.74***	−0.16*	0.41***	−0.73***
Malaysia	0.64***	0.28***	−0.64***	0.60***	0.26**	0.24**	−0.60***
Norway	1.37***	0.18**	−0.32***	0.88***	0.12**	0.90***	−0.19***
New Zealand	1.14***	0.40***	−0.84***	0.70***	0.21***	1.05***	−0.51***
Oman	1.10***	0.26*	−0.45***	0.89***	0.19***	1.12***	−0.33***
Palestinian Nat'l Auth.	0.86***	0.13	−0.35***	0.79***	0.08	0.39***	−0.36***
Qatar	0.78***	0.08	−0.67***	0.63***	−0.05	0.64***	−0.63***
Romania	1.11***	0.27***	−0.56***	0.95***	0.21***	0.43***	−0.46***
Russian Federation	1.15***	0.31***	−0.66***	0.66***	0.17***	1.15***	−0.36***
Saudi Arabia	0.96***	0.42***	−0.32**	0.66***	0.22***	1.01***	−0.24**
Singapore	0.96***	0.47***	−1.25***	0.56***	0.19***	1.18***	−0.75***
Slovenia	1.26***	0.19**	−0.24***	0.78***	0.12*	0.98***	−0.13**
Sweden	1.32***	0.12**	−0.41***	0.67***	0.05	1.16***	−0.21***
Syrian Arab Republic	0.46***	0.19**	−0.11	0.39***	0.16*	0.43***	−0.09
Thailand	0.52***	−0.04	−0.55***	0.43***	−0.08	0.55***	−0.50***
Tunisia	0.75***	0.13*	−0.32***	0.58***	0.11*	0.98***	−0.23***
Turkey	1.09***	0.17***	−0.59***	0.84***	0.08	0.67***	−0.47***
Chinese Taipei	1.31***	0.39***	−0.26***	0.58***	0.24***	1.31***	−0.03
Alabama, US	0.93***	0.15	−0.81***	0.50***	0.01	1.12***	−0.48***
California, US	0.93***	0.42***	−0.68***	0.52***	0.22***	1.13***	−0.36***
Colorado, US	1.03***	0.75***	−0.57***	0.61***	0.48***	1.11***	−0.24**
Connecticut, US	0.89***	0.67***	−0.71***	0.44***	0.41***	1.30***	−0.24**
Florida, US	0.90***	0.72***	−0.50***	0.56***	0.50***	1.10***	−0.22
Indiana, US	0.85***	0.73***	−0.35***	0.44***	0.45***	1.22***	−0.08
Ukraine	1.09***	0.26***	−0.36***	0.78***	0.13*	0.81***	−0.27***
Massachusetts, US	0.98***	0.57***	−0.54***	0.49***	0.30***	1.23***	−0.20*
Minnesota, US	1.01***	0.91***	−0.67***	0.55***	0.55***	1.19***	−0.22
North Carolina, US	0.82***	0.53***	−0.61***	0.39***	0.38***	1.30***	−0.14
United States	0.92***	0.54***	−0.63***	0.50***	0.31***	1.21***	−0.30***
South Africa (9th grade)	0.62***	−0.43***	−1.33***	0.54***	−0.49***	0.35***	−1.24***

Countries are arranged in alphabetic order of country code, but country code is not shown in the table. MSC = mathematics self-concept; BFLPE = big-fish-little-pond effect; PRS = perceived relative standing.

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

(Abu Dhabi - UAE, and Macedonia), the statistically nonsignificant effect of mathematics achievement at the between level in Model 2 became negative in Model 3.

4. Discussion and conclusions

In this study we investigated the BFLPE for eighth grade mathematics in 59 countries/regions. The BFLPE existed in most (58 out of 59) education systems. We found that the context in which students learn has two distinct influences on students' self-concept. First, data showed a statistically significant positive effect of mathematics achievement on self-concept at the classroom level in 43 of the 59 countries/regions studied. While statistically, this is simply a between-level effect of an aggregated within-level predictor, conceptually it is close to the reflected glory effect in BFLPE research. The glory effect refers to students who are in academically rigorous schools with many high achieving students enjoying the reflected glory of being in a high status institution (Marsh, Kong, & Hau, 2000; Rutter & Maughan, 2002; Trautwein, Lüdtke, Marsh, Koller, & Baumert, 2006a). Nevertheless, the positive effect of achievement on self-concept at the between level is not exactly the reflected glory because the aggregated measure representing average classroom achievement may not be indicative of perceived status of the institution. This is a weakness inherent in the data set. Separate scales specifically asking about students' perception of institution status would be better measures to examine the reflected glory effect (e.g., Marsh et al., 2000). The second influence of the learning context on students' self-concept is the BFLPE, statistically represented in the present study as the effect of a negative difference between the between-level effect and the within-level effect of mathematics achievement on mathematics self-concept. We found this effect in all countries except for the Syrian Arab Republic, suggesting generalizability of the effect.

Another important finding was that perceived relative standing was a statistically significant positive predictor of mathematics self-concept in all 59 countries/regions.

Perhaps the finding of greatest theoretical importance was that controlling for perceived relative standing eliminated the BFLPE in six countries/regions (England, Chinese Taipei, Florida-US, Indiana-US, Minnesota-US, and North Carolina-US) and reduced the BFLPE in 49 countries/regions. Huguet et al. (2009) found that the BFLPE was eliminated after the effect of perceived relative standing on self-concept was controlled, using data collected from sixth graders in French public schools. Huguet et al. assumed that forced upward social comparison with the class underlies the BFLPE, and therefore the BFLPE should be eliminated when students' perceived relative standing is controlled. The present study and Wang (2015) found that the BFLPE was generally reduced but not eliminated after controlling for perceived relative standing. Because the present study and Wang (2015) examined students in many countries/regions during two different time periods, we might assume that reduction without elimination is the typical pattern. It is also possible that differences in item wording or differences in age (8th grade versus 6th grade) influenced the results.

In Wang (2015) and the present study, students were asked to compare themselves to their classmates. However, it is difficult to know whether students had specific comparison targets in mind or not when answering the perceived relative standing question in TIMSS. Their comparison targets may be influenced by cultural and educational factors. For example, in an education system where grades are public, students may be influenced by a desire to compete or to avoid shame. The social comparison processes that result in the BFLPE may be different in different education systems. Future BFLPE research could investigate social comparison processes among different populations.

Wang (2015) used the TIMSS database from 2007 to examine the BFLPE in 49 countries, and the present study used TIMSS 2011 data. Items measuring mathematics self-concept and perceived relative standing in both studies are the same. Wang found that including perceived relative standing in the model in general decreased the BFLPE. The BFLPE was not eliminated in that study, while in the present study it was eliminated in six countries/regions. Reasons for the difference might be sample differences or slight difference in the calculation of the variance of the total mathematics self-concept [the $\mathrm{Cov}(\mathrm{math}_{wij},\mathrm{math}_{ij}^2)$ component in Eq. (9) was assumed to be zero in Wang (2015)]. Data in the current study and in the Wang (2015) study were four years apart, and 22 out of the 59 countries/regions included in the current study were new. This study provides additional evidence of the BFLPE and the relationship between perceived relative standing and the BFLPE.

Negative between-level effects of mathematics achievement on mathematics self-concept existed in four countries: Botswana, Honduras, Indonesia, and South Africa. This means that in these countries, the average mathematics self-concept of students in higher-achieving classes was lower than the average mathematics self-concept of students in lower-achieving classes. Further research could examine this unusual pattern. The Syrian Arab Republic did not have a BFLPE; the reason is not clear.

In this study, we only included the necessary constructs (mathematics self-concept, perceived relative standing, and mathematics achievement). It is possible to include student, teacher, and school covariates (student- and school-level SES, gender, language spoken at home, instructional style, etc.), but those covariates are not consistently included in BFLPE studies. Perhaps more important for this study, we could not have used the same covariates for all countries because there were country-specific adaptations of the TIMSS 2011 survey such that variables were not standardized across countries/regions. For example, some countries included information about home language and SES, but others did not. In this large-scale investigation, we wanted to use the same items for all countries/regions. Future BFLPE could consider additional covariates.

This study is an example of using large-scale assessment data in educational research. The technical details and computer codes in the appendices may serve as a resource for those who would like to use similar large-scale data.

An implication of the study is that schools should do more to reduce negative social comparison. While social comparison is normal and can foster self-concept when it enhances perceptions of competence, it can also undermine self-concept (Trautwein, Lüdtke, Marsh, & Nagy, 2009). Research on achievement goals shows that teachers can influence the degree to which students orient toward pernicious social comparison and concern with how they appear compared to others (Urdan & Schoenfelder, 2006). Research on classroom tasks and group-work provides guidance regarding how teachers can reduce social comparison, such as by avoiding the following: normative grading, comparing student competence, giving special recognition to high achievers, valuing one type of ability, and emphasizing fixed notions of ability (Cohen & Lotan, 1995; Urdan & Schoenfelder, 2006; Yeager & Dweck, 2012). Reducing these sorts of social comparison could do much to reduce disparities in self-concept and improve student well-being.

Appendix A. Variables measured on Likert-scale and the complex data structure

A.1. Perceived relative standing and mathematics self-concept items

Mathematics self-concept was measured by three items: (a) I usually do well in mathematics; (b) Mathematics is not one of my strengths; and (c) I learn things quickly in mathematics. Perceived relative standing is measured by one item: Mathematics is more difficult for me than for many of my classmates. While Marsh et al. (2014), and Marsh et al. (2015) used this item as part of their measure of self-concept, we used it to measure perceived relative standing because it asks for social comparison. One can see that it is quite different from the other three items, listed above, which all ask for ratings that do not ask for a social comparison. In addition, in the TIMSS 2007 technical report (Martin & Preuschoff, 2008), this item had the lowest loading (0.477) of the four items (the other three loadings were 0.765, 0.653 and 0.812). While TIMSS 2011 did not include corresponding loadings in its user guide or technical report, we ran a one-factor confirmatory factor analysis (CFA) with the perceived relative standing item and the three self-concept items as indicators, using all data for the present study. The maximum likelihood estimator with robust standard errors and adjusted chi-square statistics (MLR) in Mplus was used. When the data are weighted by the HOUWGT variable, the perceived relative standing item had the lowest factor loading (0.540; the other three loadings were 0.719, 0.614, and 0.740; Model fit: RMSEA = 0.245, CFI = 0.762, SRMR = 0.063). When the data were weighted by the SENWGT variable, the perceived relative standing item again had the lowest factor loading (0.577; the other three loadings were 0.729, 0.653, and 0.744; model fit: RMSEA = 0.225, CFI = 0.804, and SRMR = 0.057). Therefore, we did not include the perceived relative standing item as part of self-concept; instead, we used it as an explicit measure of social comparison. Huguet et al. (2009) explicitly measured perceived relative standing in class, and they also used a single item, one for math and one for French. Their item was very similar to the item that we are using: students were asked to rate on a scale of 1 to 5 how good they were in Math and French compared to most of their classmates.

The three mathematics self-concept items and the one perceived relative standing item were rated on a 1 to 4 Likert-scale (1 = agree a lot, 2 = agree a little, 3 = disagree a little, 4 = disagree a lot), and positive statements were reverse coded. A higher value on the items means more mathematics self-concept or more positively perceived relative standing.

A.2. Complex data structure of TIMSS 2011

In each country, there is a hierarchical data feature in that students were nested within classes from selected schools. To incorporate this complex data structure, both schools and classes were specified as clustering variables and two-level analyses were performed (in Mplus, the type of analysis is COMPLEX TWOLEVEL, which allows two cluster variables). Further, appropriate weight variables had to be used at each level. The within-level weight is the inverse of the probability that individual i in cluster j is selected, given that cluster j is selected. The between-level weight is the inverse of the probability that cluster *j* is selected. In this study, the between level is the classroom level which involves first sampling schools and then classes in the selected schools. TIMSS also includes adjustment factors to account for non-responses. Three weight factor variables are included in TIMSS: WGTFAC1, WGTFAC2, WGTFAC3, representing the probability of selecting a school, the probability of selecting a class within a selected school, and the probability of selecting a student within a selected class, respectively. There are also three adjustment variables—WGTADJ1, WGTADJ2, and WGTADJ3—for non-responses in the three sampling stages (i.e., schools, classes, and students).

For this study, when analyses were conducted for each country/region separately, in each country/region, the within-level weight was calculated as wt1 = WGTADJ3 × WGTFAC3, and the between-level weight was wt2 = WGTADJ1 × WGTFAC1 × WGTADJ2 × WGTFAC2. This is consistent with recommendations by Rutkowski, Gonzalez, Joncas, and von Davier (2010). Besides the weights themselves, the scaling of the weight at the within and between levels is also important. Asparouhov (2006) provided a procedure for selecting scaling methods. The scaling method A in Asparouhov was used in this study because the selection mechanism of the level 1 units (i.e., students) is non-invariant. With this scaling method, the within-level weight (wt1) is scaled so that the new within weights add up to the cluster sample size. This is also the default scaling method for the within-level weight in Mplus (2008). The between-level weight (wt2) is scaled using the sample scaling method. This method ensures that the product of the within and between weights sums up to the total sample size in each country. The sampling design in Australia is slightly different where all indigenous students within selected schools at eighth grade were selected regardless of whether their classes were selected (Joncas & Foy, 2012). All indigenous students within selected schools had a class weight of one, while the other sampled students from the regularly-selected TIMSS classes had the class weight of the TIMSS class that they were representing (Pierre Foy, personal communication, October 23, 2014). There were 990 indigenous eighth-graders who participated in TIMSS. They were excluded from the analysis to ensure that between-level weights for students in the same class were the same in this study.

Appendix B. SAS codes for data preparation

```
/* After using the "CONVERT.SAS" macro provided by TIMSS, all SAS datasets for all countries are saved in one folder.*/
/* A description of "CONVERT.SAS" and how to use it is on pp.4546 of the TIMSS 2011 User Guide (Foy, Arora, & Stanco, 2013) */
/* The 990 indigenous students in Australia were removed from analysis. */
/* assign a SAS library name for this folder */
libname TIMSSG8 'C:\TIMSS2011-BFLPE\G8SASdata';
%macro MyMacro2011 (cty); /* cty is the three-letter country code */
/* create a smaller SAS dataset for each country */
data a;
set TIMSSG8.bsg & cty.m5;
  keep IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX
BSBM14A BSBM14B BSBM14C BSBM14D BSBM14E BSBM14F
BSBM16A BSBM16B BSBM16C BSBM16D BSBM16E BSBM16F BSBM16G BSBM16H BSBM16I BSBM16J BSBM16K BSBM16L BSBM16M BSBM16N
BSMMAT01 BSMMAT02 BSMMAT03 BSMMAT04 BSMMAT05
TOTWGT HOUWGT SENWGT WGTADJ1 WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2
WGTFAC3 JKZONE JKREP;
run;
/* reverse coding and code invalid values as missing */
data aa;
  set a;
  BSBM14A = 5-BSBM14A;
  BSBM14D = 5-BSBM14D;
  BSBM14E = 5-BSBM14E;
  BSBM14F = 5-BSBM14F;
  BSBM16A = 5-BSBM16A;
  BSBM16D = 5-BSBM16D;
  BSBM16F = 5-BSBM16F;
  BSBM16G = 5-BSBM16G;
  BSBM16H = 5-BSBM16H;
  BSBM16J = 5-BSBM16J;
  BSBM16K = 5-BSBM16K;
  BSBM16L = 5-BSBM16L;
  BSBM16M = 5-BSBM16M;
  BSBM16N = 5-BSBM16N;
  BSBM14B = 0 + BSBM14B;
  BSBM14C = 0 + BSBM14C;
  BSBM16B = 0 + BSBM16B;
  BSBM16C = 0 + BSBM16C;
  BSBM16E = 0 + BSBM16E;
  BSBM16I = 0 + BSBM16I;
run;
/* standardize and rename variables */
proc standard data = aa mean = 0 std = 1 out = aaa;
var BSBM16A BSBM16B BSBM16C BSBM16D BSBM16E BSBM16F BSBM16G BSBM16H BSBM16I BSMMAT01 BSMMAT02 BSMMAT03 BSMMAT04 BSMMAT05;
```

```
run; data aaa (rename = (BSBM16A = zBSBM16A BSBM16B
= zBSBM16B BSBM16C = zBSBM16C BSBM16D = zBSBM16D BSBM16E
= zBSBM16E BSBM16F = zBSBM16F BSBM16G = zBSBM16G BSBM16H
= zBSBM16H BSBM16I = zBSBM16I BSMMAT01 = zmath1 BSMMAT02
= zmath2 BSMMAT03 = zmath3 BSMMAT04 = zmath4 BSMMAT05
= zmath5));
    set aaa;
run;
/* replace missing values on standardized variables with 9's */
data aaa;
    set aaa;
    if zBSBM16A = . then zBSBM16A = 9;
    if zBSBM16B = . then zBSBM16B = 9;
    if zBSBM16C = . then zBSBM16C = 9;
    if zBSBM16D = . then zBSBM16D = 9;
    if zBSBM16E = . then zBSBM16E = 9;
    if zBSBM16F = . then zBSBM16F = 9;
    if zBSBM16G = . then zBSBM16G = 9;
    if zBSBM16H = . then zBSBM16H = 9;
    if zBSBM16I = . then zBSBM16I = 9;
    if ((ITSEX ~ = 1) & (ITSEX ~ = 2)) then ITSEX = 9;
run;
/* create five datasets for each country. Each dataset has one
plausible value for math achievement */
data z & cty.math1 (drop = zmath2 zmath3 zmath4 zmath5 re-
name = (zmath1 = zmath))
    z & cty.math2 (drop = zmath1 zmath3 zmath4 zmath5 rename =
(zmath2 = zmath))
z & cty.math3 (drop = zmath1 zmath2 zmath4 zmath5 rename =
(zmath3 = zmath))
    z & cty.math4 (drop = zmath1 zmath2 zmath3 zmath5 rename =
(zmath4 = zmath))
    z & cty.math5 (drop = zmath1 zmath2 zmath3 zmath4 rename =
(zmath5 = zmath));
    set aaa;
run;
/* create five .dat files for each country. Each .dat file has one
plausible value for math achievement */
data _null_;
    set z & cty.math1;
file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math1.dat";
put IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX zBSBM16A
zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F zBSBM16G
zBSBM16H zBSBM16I zmath
TOTWGT HOUWGT SENWGT WGTADJ1 WGTADJ2 WGTADJ3
WGTFAC1 WGTFAC2
WGTFAC3 JKZONE JKREP;
run;
data _null_;
    set z & cty.math2;
    file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math2.dat";
    put IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX zBSBM16A
zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F zBSBM16G
zBSBM16H zBSBM16I zmath TOTWGT HOUWGT SENWGT WGTADJ1
WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2
WGTFAC3 JKZONE JKREP;
run;
data _null_;
    set z & cty.math3;
    file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math3.dat";
    put IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX zBSBM16A
zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F zBSBM16G
zBSBM16H zBSBM16I zmath TOTWGT HOUWGT SENWGT WGTADJ1
WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2 WGTFAC3 JKZONE
JKREP;
run;
data _null_;
    set z & cty.math4;
    file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math4.dat";
put IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX zBSBM16A
zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F zBSBM16G
zBSBM16H zBSBM16I zmath TOTWGT HOUWGT SENWGT WGTADJ1
WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2 WGTFAC3 JKZONE
JKREP;
run;
data _null_;
    set z & cty.math5;
    file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math5.dat";
put IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX zBSBM16A
zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F zBSBM16G
zBSBM16H zBSBM16I zmath TOTWGT HOUWGT SENWGT WGTADJ1
WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2 WGTFAC3 JKZONE
JKREP;
run;
/* create a dataset with a list of the five datasets for each country */
data _null_;
    file "C:\TIMSS2011-BFLPE\batch runs in R\z & cty.math.dat";
    put "z & cty.math1.dat";
    put "z & cty.math2.dat";
    put "z & cty.math3.dat";
    put "z & cty.math4.dat";
    put "z & cty.math5.dat";
run;
%mend MyMacro2011;
/* apply macro to 59 countries/regions */
%MyMacro2011(ARM);
%MyMacro2011(AUS);
%MyMacro2011(BHR);
%MyMacro2011(CHL);
%MyMacro2011(TWN);
%MyMacro2011(ENG);
%MyMacro2011(FIN);
%MyMacro2011(GEO);
%MyMacro2011(GHA);
%MyMacro2011(HKG);
%MyMacro2011(HUN);
%MyMacro2011(IDN);
%MyMacro2011(IRN);
%MyMacro2011(ISR);
%MyMacro2011(ITA);
%MyMacro2011(JPN);
%MyMacro2011(JOR);
%MyMacro2011(KAZ);
%MyMacro2011(KOR);
%MyMacro2011(LBN);
%MyMacro2011(LTU);
%MyMacro2011(MYS);
%MyMacro2011(MKD);
%MyMacro2011(MAR);
%MyMacro2011(NZL);
%MyMacro2011(NOR);
%MyMacro2011(OMN);
%MyMacro2011(PSE);
%MyMacro2011(QAT);
%MyMacro2011(ROM);
%MyMacro2011(RUS);
%MyMacro2011(SAU);
%MyMacro2011(SGP);
%MyMacro2011(SVN);
%MyMacro2011(SWE);
%MyMacro2011(SYR);
%MyMacro2011(THA);
```

```
%MyMacro2011(TUN);
%MyMacro2011(TUR);
%MyMacro2011(UKR);
%MyMacro2011(ARE);
%MyMacro2011(USA);
%MyMacro2011(BWA);
%MyMacro2011(HND);
%MyMacro2011(ZAF);
%MyMacro2011(CAB);
%MyMacro2011(COT);
%MyMacro2011(CQU);
%MyMacro2011(AAD);
%MyMacro2011(ADU);
%MyMacro2011(UAL);
%MyMacro2011(UCA);
%MyMacro2011(UCO);
%MyMacro2011(UCT);
%MyMacro2011(UFL);
%MyMacro2011(UIN);
%MyMacro2011(UMA);
%MyMacro2011(UMN);
%MyMacro2011(UNC);
run;
/* merge all data in 59 countries/regions for descriptive statistics */
data TIMSSG8.alldata alldata;
  set TIMSSG8.bsgarmm5
  TIMSSG8.bsgausm5
  TIMSSG8.bsgbhrm5
  TIMSSG8.bsgchlm5
  TIMSSG8.bsgtwnm5
  TIMSSG8.bsgengm5
  TIMSSG8.bsgfinm5
  TIMSSG8.bsggeom5
  TIMSSG8.bsggham5
  TIMSSG8.bsghkgm5
  TIMSSG8.bsghunm5
  TIMSSG8.bsgidnm5
  TIMSSG8.bsgirnm5
  TIMSSG8.bsgisrm5
  TIMSSG8.bsgitam5
  TIMSSG8.bsgjpnm5
  TIMSSG8.bsgjorm5
  TIMSSG8.bsgkazm5
  TIMSSG8.bsgkorm5
  TIMSSG8.bsglbnm5
  TIMSSG8.bsgltum5
  TIMSSG8.bsgmysm5
  TIMSSG8.bsgmkdm5
  TIMSSG8.bsgmarm5
  TIMSSG8.bsgnzlm5
  TIMSSG8.bsgnorm5
  TIMSSG8.bsgomnm5
  TIMSSG8.bsgpsem5
  TIMSSG8.bsgqatm5
  TIMSSG8.bsgromm5
  TIMSSG8.bsgrusm5
  TIMSSG8.bsgsaum5
  TIMSSG8.bsgsgpm5
  TIMSSG8.bsgsvnm5
  TIMSSG8.bsgswem5
  TIMSSG8.bsgsyrm5
  TIMSSG8.bsgtham5
  TIMSSG8.bsgtunm5
  TIMSSG8.bsgturm5
  TIMSSG8.bsgukrm5
  TIMSSG8.bsgarem5
  TIMSSG8.bsgusam5
  TIMSSG8.bsgbwam5
  TIMSSG8.bsghndm5
  TIMSSG8.bsgzafm5
  TIMSSG8.bsgcabm5
  TIMSSG8.bsgcotm5
  TIMSSG8.bsgcqum5
  TIMSSG8.bsgaadm5
  TIMSSG8.bsgadum5
  TIMSSG8.bsgualm5
  TIMSSG8.bsgucam5
  TIMSSG8.bsgucom5
  TIMSSG8.bsguctm5
  TIMSSG8.bsguflm5
  TIMSSG8.bsguinm5
  TIMSSG8.bsgumam5
  TIMSSG8.bsgumnm5
  TIMSSG8.bsguncm5;
run;
proc sort data = alldata out = sortalldata;
  by idcntry idschool idclass;
run;
/* count number of schools by country/region */
data schools;
  set sortalldata;
  by idcntry idschool;
  retain NumSchools;
  if (first.idcntry & first.idschool) then NumSchools = 1;
  else if (first.idcntry = 0 & first.idschool)then NumSchools
= NumSchools + 1;
  if last.idcntry then output;
  keep idcntry NumSchools;
run;
/* count number of classes by country/region */
data classes;
  set sortalldata;
  by idcntry idclass;
  retain NumClasses;
  if (first.idcntry & first.idclass) then NumClasses = 1;
  else if (first.idcntry = 0 & first.idclass)then NumClasses
= NumClasses + 1;
  if last.idcntry then output;
  keep idcntry NumClasses;
run;
/* count number of students by country/region */
data students;
  set sortalldata;
  by idcntry;
  retain NumStudents;
  if first.idcntry then NumStudents = 0;
  NumStudents = NumStudents + 1;
  if last.idcntry then output;
  keep idcntry NumStudents;
run;
data TIMSSG8.NumByCntry NumByCntry;
  merge schools classes students;
  by idcntry;
run;
/* export descriptive statistics by country to Excel */
PROC EXPORT DATA = TIMSSG8.NumByCntry
OUTFILE = "C:\TIMSS2011-BFLPE\NumbersbyCountry.xlsx"
  DBMS = EXCEL REPLACE;
SHEET = "NumbersbyCountry";
RUN;
```

Appendix C. Mplus templates

C.1. Mplus template for Model 1

```
! This is an Mplus template file. This template is used to create a
! group of Mplus input files, consistent with Statistical Model 1
! in manuscript, for the 59 countries/regions.
! The init section is used to specify the number of iterations to loop
! over, the filenames and directories for the input files created by
! createModels of the R "MplusAutomation" package, and the fields to be
! inserted in the body section where template tags are specified.
[[init]]
iterators = datafile; ! loop over a variable called datafile
datafile = 1:59;
datafileNames#datafile =
zARMmath1 zBHRmath1 zCHLmath1 zTWNmath1 zENGmath1 zFINmath1 zGEOmath1
zGHAmath1 zHKGmath1 zHUNmath1 zIDNmath1 zIRNmath1 zISRmath1 zITAmath1
zJPNmath1 zJORmath1 zKAZmath1 zKORmath1 zLBNmath1 zLTUmath1 zMYSmath1
zMKDmath1 zMARmath1 zNZLmath1 zNORmath1 zOMNmath1 zPSEmath1 zQATmath1
zROMmath1 zRUSmath1 zSAUmath1 zSGPmath1 zSVNmath1 zSWEmath1 zSYRmath1
zTHAmath1 zTUNmath1 zTURmath1 zUKRmath1 zAREmath1 zUSAmath1 zBWAmath1
zHNDmath1 zZAFmath1 zCABmath1 zCOTmath1 zCQUmath1 zAADmath1 zADUmath1
zUALmath1 zUCAmath1 zUCOmath1 zUCTmath1 zUFLmath1 zUINmath1 zUMAmath1
zUMNmath1 zUNCmath1 zAUSmath1;! a total of 59 datafiles
! Model 1 does not involve math achievement, datasets with the first
! plausible value are used.
filename = "TIMSS2011Model1[[datafileNames#datafile]].inp";
outputDirectory = "C:/TIMSS2011-BFLPE/batch runs in R/Model 1";
[[/init]]
TITLE: Model1
DATA: FILE IS "C:/TIMSS2011-BFLPE/batch runs in R/
[[datafileNames#datafile]].dat";
VARIABLE: NAMES ARE IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX
zBSBM16A zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F
zBSBM16G zBSBM16H zBSBM16I zmath TOTWGT HOUWGT SENWGT
WGTADJ1 WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2 WGTFAC3 JKZONE JKREP;
! wgtfac1 and wgtadj1 are school weighting factor and adjustment
! wgtfac2 and wgtadj2 are class weighting factor and adjustment
! wgtfac3 and wgtadj3 are student weighting factor and adjustment
USEVARIABLES ARE IDSCHOOL IDCLASS zBSBM16A zBSBM16C zBSBM16D wt1 wt2;
! zBSBM16A - zBSBM16D are math self-concept items, standardized
! within each country/region
MISSING ARE ITSEX(9) zBSBM16A-zBSBM16I(9) TOTWGT-WGTFAC3(999999.000000) JKZONE(99) JKREP(9); ! specify missing values for variables
CLUSTER = IDSCHOOL IDCLASS; ! clustering within classes and schools
WEIGHT IS wt1; ! within-level weight
WTSCALE IS CLUSTER;
! CLUSTER is default; it rescales within level weights so that
! they sum to cluster size;
BWEIGHT IS wt2; ! between-level weight;
BWTSCALE IS SAMPLE;
! SAMPLE is default; it adjusts the between weights so that the
! products of the between and within weights sum to the total sample size.
DEFINE: wt1 = WGTADJ3*WGTFAC3;
wt2 = WGTADJ1*WGTFAC1*WGTADJ2*WGTFAC2;
zmathsq = zmath*zmath;
ANALYSIS: TYPE = COMPLEX TWOLEVEL;
MODEL: %WITHIN%
scw BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);! within-level CFA
scw* (var1);! within-level variance of math self-concept
%BETWEEN%
scb BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);! between-level CFA
scb* (var2);! between-level variance of math self-concept
OUTPUT: STANDARDIZED;
```

C.2. Mplus template for Model 2

```
! This is an Mplus template file. This template is used to create a
! group of Mplus input files, consistent with Statistical Model 2
! in manuscript, for the 59 countries/regions.
! The init section is used to specify the number of iterations to loop
! over, the filenames and directories for the input files created by
! createModels of the R "MplusAutomation" package, and the fields to be
! inserted in the body section where template tags are specified.
[[init]]
iterators = datafile; ! loop over a variable called datafile
datafile = 1:59;
datafileNames#datafile =
zARMmath zBHRmath zCHLmath zTWNmath zENGmath zFINmath zGEOmath zGHAmath zHKGmath zHUNmath zIDNmath zIRNmath zISRmath zITAmath zJPNmath zJORmath
zKAZmath zKORmath zLBNmath zLTUmath zMYSmath zMKDmath zMARmath zNZLmath
zNORmath zOMNmath zPSEmath zQATmath zROMmath zRUSmath zSAUmath zSGPmath
zSVNmath zSWEmath zSYRmath zTHAmath zTUNmath zTURmath zUKRmath zAREmath
zUSAmath zBWAmath zHNDmath zZAFmath zCABmath zCOTmath zCQUmath zAADmath
zADUmath zUALmath zUCAmath zUCOmath zUCTmath zUFLmath zUINmath zUMAmath
zUMNmath zUNCmath zAUSmath;
! a total of 59 datafiles
! five datasets' names are in each of the above file for each country.
filename = "TIMSS2011Model2[[datafileNames#datafile]].inp";
outputDirectory = "C:/TIMSS2011-BFLPE/batch runs in R/Model 2";
[[/init]]
TITLE: Model2
DATA: FILE IS "C:/TIMSS2011-BFLPE/batch runs in R/
[[datafileNames#datafile]].dat";
! The file for each country/region has the names of five datasets.
! Those five datasets include variable values. For the math
! achievement variable, the five series of plausible values are
! included in the five datasets respectively.
TYPE IS IMPUTATION;
! The five datasets for each country/region are imputed datasets,
! consistent with the way the plausible values were created in
! TIMSS.
```

```
VARIABLE: NAMES ARE IDCNTRY IDSCHOOL IDCLASS IDSTUD ITSEX
zBSBM16A zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F
zBSBM16G zBSBM16H zBSBM16I zmath
TOTWGT HOUWGT SENWGT WGTADJ1 WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2
WGTFAC3 JKZONE JKREP;
! wgtfac1 and wgtadj1 are school weighting factor and adjustment
! wgtfac2 and wgtadj2 are class weighting factor and adjustment
! wgtfac3 and wgtadj3 are student weighting factor and adjustment
USEVARIABLES ARE IDSCHOOL IDCLASS zBSBM16A zBSBM16C zBSBM16D zmath wt1 wt2 zmathsq;
! zBSBM16A – zBSBM16D are math self-concept items, standardized within
! each country/region. zmath is the math achievement standardized within
! each series of plausible values in each country/region.
MISSING ARE ITSEX(9) zBSBM16A-zBSBM16I(9) TOTWGT-WGTFAC3(999999.000000)
  JKZONE(99) JKREP(9);! specify missing values for variables
WITHIN = zmathsq;
! quadratic term of math achievement as a within-level variable
CLUSTER = IDSCHOOL IDCLASS; ! clustering within classes and schools
WEIGHT IS wt1; ! within-level weight;
WTSCALE IS CLUSTER;
BWEIGHT IS wt2; ! between-level weight;
BWTSCALE IS SAMPLE;
DEFINE: wt1 = WGTADJ3*WGTFAC3;
wt2 = WGTADJ1*WGTFAC1*WGTADJ2*WGTFAC2;
zmathsq = zmath*zmath;
ANALYSIS: TYPE = COMPLEX TWOLEVEL;
MODEL: %WITHIN%
scw BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);
scw ON zmathsq (bb)
zmath (b1);
! the linear and quadratic terms of math achievement are predictors of
! math self-concept at within level
scw (var1);
zmath (var2);
zmathsq (var5);
zmath with zmathsq (cov1);
%BETWEEN%
scb BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);
scb on
zmath (b2);
! math achievement as predictor of math self-concept at between level
scb (var3);
zmath (var4);
MODEL CONSTRAINT:
NEW(ESw);
NEW(ESb);
NEW(ESBFLPE);
ESw = 2*b1*sqrt(var2)/sqrt(b1**2*var2 + bb**2*-var5 + 2*b1*bb*cov1
 + var1 + b2**2*var4 + var3); ! equation (10) in manuscript.
ESb = 2*b2*sqrt(var4)/sqrt(b1**2*var2 + bb**2*var5 + 2*b1*bb*-cov1 + var1
 + b2**2*var4 + var3); ! equation (11) in manuscript.
ESBFLPE = 2*(b2-b1)*sqrt(var4)
/sqrt(b1**2*var2 + bb**2*var5 + 2*b1*bb*cov1 + var1 + b2**2*-var4 + var3);
! ESw is the size of within-level linear effect of math ability on sc.
! ESb is the size of between-level effect of math ability on sc.
! ESBFLPE is the size of contextual effect (i.e., BELPE).
OUTPUT: STANDARDIZED;
```

C.3. Mplus Template for Model 3

```
! This is an Mplus template file. This template is used to create a
! group of Mplus input files, consistent with Statistical Model 3
! in manuscript, for the 59 countries/regions
! The init section is used to specify the number of iterations to loop
! over, the filenames and directories for the input files created by
! createModels of the R "MplusAutomation" package, and the fields to be
! inserted in the body section where template tags are specified
[[init]]
iterators = datafile; ! loop over a variable called datafile
datafile = 1:59;
datafileNames#datafile =
zAUSmath zBHRmath zCHLmath zTWNmath zENGmath zFINmath zGEOmath zGHAmath
zHKGmath zHUNmath zIDNmath zIRNmath zISRmath zITAmath zJPNmath zJORmath
zKAZmath zKORmath zLBNmath zLTUmath zMYSmath zMKDmath zMARmath zNZLmath
zNORmath zOMNmath zPSEmath zQATmath zROMmath zRUSmath zSAUmath zSGPmath
zSVNmath zSWEmath zSYRmath zTHAmath zTUNmath zTURmath zUKRmath zAREmath
zUSAmath zBWAmath zHNDmath zZAFmath zCABmath zCOTmath zCQUmath zAADmath
zADUmath zUALmath zUCAmath zUCOmath zUCTmath zUFLmath zUINmath zUMAmath
zUMNmath zUNCmath zAUSmath;
! a total of 59 datafiles
! five datasets' names are in each of the above file for each country
filename = "TIMSS2011Model3[[datafileNames#datafile]].inp";
outputDirectory = "C:/TIMSS2011-BFLPE/batch runs in R/Model 3";
[[/init]]
TITLE: Model3
DATA: FILE IS "C:/TIMSS2011-BFLPE/batch runs in R/
[[datafileNames#datafile]].dat";
! The file for each country/region has the names of five datasets
! Those five datasets include variable values. For the math
! achievement variable, the five series of plausible values are
! included in the five datasets respectively.
TYPE IS IMPUTATION;
! The five datasets for each country/region are imputed datasets,
! consistent with the way the plausible values were created in.
TIMSS.
VARIABLE: NAMES ARE JKREP IDSCHOOL IDCLASS IDSTUD ITSEX
zBSBM16A zBSBM16B zBSBM16C zBSBM16D zBSBM16E zBSBM16F
zBSBM16G zBSBM16H zBSBM16I zmath
TOTWGT HOUWGT SENWGT WGTADJ1 WGTADJ2 WGTADJ3 WGTFAC1 WGTFAC2
WGTFAC3 JKZONE JKREP;
! wgtfac1 and wgtadj1 are school weighting factor and adjustment
! wgtfac2 and wgtadj2 are class weighting factor and adjustment
! wgtfac3 and wgtadj3 are student weighting factor and adjustment
USEVARIABLES ARE IDSCHOOL IDCLASS zBSBM16A zBSBM16B zBSBM16C zBSBM16D zmath wt1 wt2 zmathsq;
```

```
! zBSBM16A – zBSBM16D are math self-concept items, standardized within
each country/region. zmath is the math achievement standardized within
! each series of plausible values in each country/region.
MISSING ARE ITSEX(9) zBSBM16A-zBSBM16I(9) TOTWGT-WGTFAC3(999999.000000)
JKZONE(99) JKREP(9); ! specify missing values for variables
WITHIN = zmathsq zBSBM16B;
CLUSTER = IDSCHOOL IDCLASS;! clustering within classes and schools
WEIGHT IS wt1; ! within-level weight;
WTSCALE IS CLUSTER;
BWEIGHT IS wt2; ! between-level weight;
BWTSCALE IS SAMPLE;
DEFINE: wt1 = WGTADJ3*WGTFAC3;
wt2 = WGTADJ1*WGTFAC1*WGTADJ2*WGTFAC2;
zmathsq = zmath*zmath;
ANALYSIS: TYPE = COMPLEX TWOLEVEL;
MODEL: %WITHIN%
scw BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);
scw ON zBSBM16B (b3)
zmathsq (bb)
zmath (b1);
! perceived relative standing, and the linear and quadratic terms of
! math achievement are predictors of math self-concept at within level
scw (var1);
zmath (var2);
zmathsq (var5);
zBSBM16B (var6);
zmath with zmathsq (cov1);
zmath with zBSBM16B (cov2);
zmathsq with zBSBM16B (cov3);
%BETWEEN%
scb BY zBSBM16A (1)
zBSBM16C (3)
zBSBM16D (4);
scb on
zmath (b2);
! math achievement as predictor of math self-concept at between
level
scb (var3);
zmath (var4);
MODEL CONSTRAINT:
NEW(ESw);
NEW(ESb);
NEW(ESprs);
NEW(ESBFLPE);
ESw = 2*b1*sqrt(var2)/sqrt(b1**2*var2 + bb**2*var5 + b3**2*var6 +      2*b1*bb*cov1 + 2*b1*b3*cov2 + 2*bb*b3*cov3 + var1 + b2**2*var4 + var3);
ESb = 2*b2*sqrt(var4)/sqrt(b1**2*var2 + bb**2*var5 + b3**2*var6 +      2*b1*bb*cov1 + 2*b1*b3*cov2 + 2*bb*b3*cov3 + var1 + b2**2*var4 + var3);
ESprs = 2*b3*sqrt(var6)/sqrt(b1**2*var2 + bb**2*var5 + b3**2*var6 +      2*b1*bb*cov1 + 2*b1*b3*cov2 + 2*bb*b3*cov3 + var1 + b2**2*var4 + var3);
ESBFLPE = 2*(b2-b1)*sqrt(var4)/sqrt(b1**2*var2 + bb**2*var5 + b3**2*var6 +      2*b1*bb*cov1 + 2*b1*b3*cov2 + 2*bb*b3*cov3 + var1 + b2**2*var4 + var3);
! ESw is the size of within-level linear effect of math ability on sc;
! ESb is the size of between-level effect of math ability on sc;
! ESprs is the size of perceived relative standing on sc.
! This is within-level effect standardized by total sc variance;
! ESBFLPE is the size of contextual effect (i.e., BELPE);
OUTPUT: STANDARDIZED;
```

Appendix D. R codes of using package "MplusAutomation" to create, run, and extract results

```
install.package("MplusAutomation", .Library)
library(MplusAutomation)
citation(package = "MplusAutomation")
## create models using Mplus templates and run these models by calling Mplus in R ##
> createModels("C:/TIMSS2011-BFLPE/batch runs in R/TIMSS2011syntaxtemplateModel1.inp")
> runModels("C:/TIMSS2011-BFLPE/batch runs in R/Model 1")
> createModels("C:/TIMSS2011-BFLPE/batch runs in R/TIMSS2011syntaxtemplateModel2.inp")
> runModels("C: TIMSS2011-BFLPE/batch runs in R/Model 2")
> createModels("C:/TIMSS2011-BFLPE/batch runs in R/TIMSS2011syntaxtemplateModel3.inp")
> runModels("C:/TIMSS2011-BFLPE/batch runs in R/Model 3")
## extract model summaries and model parameters ##
> model1Summaries < -extractModelSummaries("C:/TIMSS2011-BFLPE/batch runs in R/Model 1", filefilter = "timss2011model1.*-math1")
> model1Summaries
> showSummaryTable(model1Summaries,keepCols = c("Filename", "Parameters", "ChiSqM_Value", "ChiSqM_DF", "ChiSqM_PValue", "CFI", "TLI", "AIC", "BIC", "RMSEA_Estimate", "SRMR.Within", "SRMR.Between"),sortBy = "Filename")
> HTMLSummaryTable(model1Summaries,filename = "C:/Users/wangze/Desktop/TIMSS2011-BFLPE/batch runs in R/Model 1/Model1Summary.html", display = TRUE, keepCols = c("Filename", "Parameters", "ChiSqM_Value", "ChiSqM_DF", "ChiSqM_PValue", "CFI", "TLI", "AIC", "BIC", "RMSEA_Estimate", "SRMR.Within", "SRMR.Between"),sortBy = "Filename")
> model2Summaries < -extractModelSummaries("C:/TIMSS2011-BFLPE/batch runs in R/Model 2", filefilter = "timss2011model2.*math")
> model2Summaries
> model2Parameters < -extractModelParameters("C:/TIMSS2011-BFLPE/batch runs in R/Model 2", filefilter = "timss2011model2.*math")
> Model2Unstandardized < -sapply(model2Parameters, "[", "unstandardized")
> Model2Unstandardized
> lapply(names(Model2Unstandardized), function(element){
Model2Unstandardized[[element]]$filename < < -element})
> Model2combinedParameters < -do.call("rbind", Model2Unstandardized)
> Model2combinedParameters
> model3Summaries < -extractModelSummaries("C:/TIMSS2011-BFLPE/batch runs in R/Model 3", filefilter = "timss2011model3.*math")
model3Summaries
> model3Parameters < -extractModelParameters("C:/TIMSS2011-BFLPE/batch runs in R/Model 3", filefilter = "timss2011model3.*math")
> Model3Unstandardized < -sapply(model3Parameters, "[", "unstandardized")
> Model3Unstandardized
> lapply(names(Model3Unstandardized), function(element){
Model3Unstandardized[[element]]$filename < < -element})
> Model3combinedParameters < -do.call("rbind", Model3Unstandardized)
> Model3combinedParameters
##process model 2 parameters for Table 2 in paper ##
> nrow(model2Summaries)
> nrow(Model2combinedParameters)
# there are 29 parameters in each model: 1711/59 = 29
```

```
> Model2Unstandardized
#ESw is the 27th parmater; i1
#ESb is the 28th parameter; i2
#ESBFLPE is the 29th parameter; i3
> rm(m, i1, i2, i3)
> m < -1:59
> i1 < -29*m-2
> i2 < -29*m-1
> i3 < -29*m
> model2ESw < -Model2combinedParameters[i1,]
> model2ESb < -Model2combinedParameters[i2,]
> model2ESBFLPE < -Model2combinedParameters[i3,]
> names(model2ESw)
> Table2_Model2 < -data.frame(model2ESw$est, model2ESw
$pval,model2ESb$est, model2ESb$pval,model2ESBFLPE$est, model
2ESBFLPE$pval )
> Table2_Model2
## process model 3 parameters for Table 2 in paper ##
> nrow(model3Summaries)
> nrow(Model3combinedParameters)
# there are 32 parameters in each model: 2065/59 = 35
> Model3Unstandardized
#ESw is the 32nd parmater; n1
#ESb is the 33th parameter; n2
#ESprs is the 34th parameter; n3
#ESBFLPE is the 35th parameter; n4
> rm(m, n1, n2, n3, n4)
> m < -1:59
> n1 < -35*m-3
> n2 < -35*m-2
> n3 < -35*m-1
> n4 < -35*m
> model3ESw < -Model3combinedParameters[n1,]
> model3ESb < -Model3combinedParameters[n2,]
> model3ESprs < -Model3combinedParameters[n3,]
> model3ESBFLPE < -Model3combinedParameters[n4,]
> names(model3ESw)
> Table2_Model3 < -data.frame(model3ESw$est, model3ESw
$pval, model3ESb$est, model3ESb$pval,model3ESprs$est, model3
ESprs$pval, model3ESBFLPE$est, model3ESBFLPE$pval )
> Table2_Model3
```

References

Arens, A. K., & Watermann, R. (2015). How an early transition to high-ability secondary schools affects students' academic self-concept: Contrast effects, assimilation effects, and differential stability. *Learning and Individual Differences*, *37*, 64–71. http://dx.doi.org/10.1016/j.lindif.2014.11.007.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, *35*, 439–460.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

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

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

Cohen, E. G., & Lotan, R. A. (1995). Producing equal-status interaction in the heterogeneous classroom. *American Educational Research Journal*, *32*, 99–120. http://dx.doi.org/10.3102/00028312032001099.

Dai, D., & Rinn, A. (2008). The big-fish-little-pond effect: What do we know and where do we go from here? *Educational Psychology Review*, *20*, 283–317. http://dx.doi.org/10.1007/s10648-008-9071-x.

In Foy, P., Arora, A., & Stanco, G. M. (Eds.), (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill: MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, and International Association for the Evaluation of Educational Achievement (IEA).

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, *95*, 124–136. http://dx.doi.org/10.1037/0022-0663.95.1.124.

Hallquist, M. (2011). MplusAutomation: Automating Mplus model estimation and interpretation. (R package version 0.5) [online]. Retrieved from http://CRAN.R-project.org/package=MplusAutomation.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. http://dx.doi.org/10.1037/1082-989x.3.4.424.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., et al. (2009). Clarifying the role of social comparison in the big-fish–little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*, 156–170. http://dx.doi.org/10.1037/a0015558.

Ireson, J., & Hallam, S. (2005). Pupils' liking for school: Ability grouping, self-concept and perceptions of teaching. *British Journal of Educational Psychology*, *75*, 297–311. http://dx.doi.org/10.1348/000709904X24762.

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2015). Contrast and assimilation effects of dimensional comparisons in five subjects: An extension of the I/E model. *Journal of Educational Psychology*, *107*, 1086–1101. http://dx.doi.org/10.1037/edu000002110.1037/edu0000021 (supp (Supplemental)).

In Joncas, M., & Foy, P. (Eds.), (2012). *Sample design in TIMSS and PIRLS*. Boston, MA: TIMSS & PIRLS International Study Center.

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal*, *50*, 326–370. http://dx.doi.org/10.3102/0002831212464511.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832. http://dx.doi.org/10.1177/0956797615616374.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. http://dx.doi.org/10.1037/a0012869.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*, 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280.

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

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, *81*, 59–77. http://dx.doi.org/10.1348/000709910X503501.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*, 213–231. http://dx.doi.org/10.1037/0022-3514.47.1.213.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*, 337–349. http://dx.doi.org/10.1037/0022-3514.78.2.337.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K., O'Mara, A. J., et al. (2008a). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*, 319–350. http://dx.doi.org/10.1007/s10648-008-9075-6.

Marsh, H. W., Trautwein, U., Lüdtke, O., & Köller, O. (2008b). Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others. *Journal of Educational Psychology*, *100*(3), 510–524. http://dx.doi.org/10.1037/0022-0663.100.3.510.

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

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

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., & Nagengast, B. (2014). The big-fish-little-pond effect in mathematics: A cross-cultural comparison of U.S. and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology*, *45*, 777–804. http://dx.doi.org/10.1177/0022022113519858.

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P., Abdelfattah, F., Nagengast, B., et al. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology*, *107*, 258–271. http://dx.doi.org/10.1037/a0037485.

Martin, M. O., & Preuschoff, C. (2008). Creating the TIMSS 2007 background indices. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 281–338). Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mplus (2008). *Scaling of sampling weights for two level models in Mplus 4.2. Technical appendix*.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*(7th ed.). Los Angeles, CA: Muthén & Muthén.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), http://dx.doi.org/10.1126/science.aac4716 (aac4716).

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142–151. http://dx.doi.org/10.3102/0013189x10363170.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology*, *40*, 451–475. http://dx.doi.org/10.1016/S0022-4405(02)00124-3.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441.

Swann, W. B., Chang-Schneider, C., & McClarty, K. L. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist*, *62*, 84–94. http://dx.doi.org/10.1037/0003-066X.62.2.84.

Trautwein, U., Lüdtke, O., Marsh, H. W., Koller, O., & Baumert, J. (2006a). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*, 788–806. http://dx.doi.org/10.1037/0022-0663.98.4.788.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006b). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology*, *98*, 438–456. http://dx.doi.org/10.1037/0022-0663.98.2.438.

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

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology*, *44*, 331–349. http://dx.doi.org/10.1016/j.jsp.2006.04.003.

Wang, Z. (2015). Examining big-fish-little-pond-effects across 49 countries: A multilevel latent variable modelling approach. *Educational Psychology*, *35*, 228–251. http://dx.doi.org/10.1080/01443410.2013.827155.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*, 302–314. http://dx.doi.org/10.1080/00461520.2012.722805.