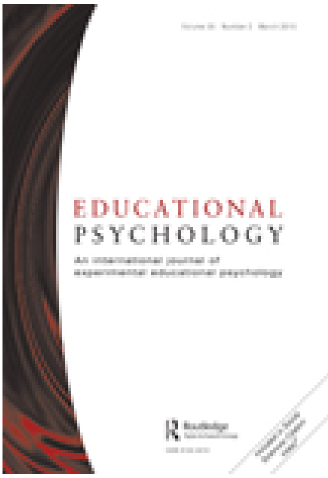

Educational Psychology: An International Journal of Experimental Educational Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cedp20

Examining big-fish-little-pond-effects across 49 countries: a multilevel latent variable modelling approach

Ze Wang[a]
[a] Educational, School and Counseling Psychology, University of Missouri, Columbia, MO, USA
Published online: 12 Aug 2013.

Educational Psychology, 2015
Vol. 35, No. 2, 228–251, http://dx.doi.org/10.1080/01443410.2013.827155

Examining big-fish-little-pond-effects across 49 countries: a multilevel latent variable modelling approach

Ze Wang*

Educational, School and Counseling Psychology, University of Missouri, Columbia, MO, USA

(*Received 12 October 2012; final version received 5 July 2013*)

Using data from the Trends in International Mathematics and Science Study (TIMSS) 2007, this study examined the big-fish-little-pond-effects (BFLPEs) in 49 countries. In this study, the effect of math ability on math self-concept was decomposed into a within- and a between-level components using implicit mean centring and the complex data structure was considered. Sampling error was thus reduced. In addition, measurement error was minimised by modelling math self-concept as a latent factor at both levels and multilevel measurement invariance was assessed in each country. The BFLPE in each country was a contextual effect modelled as the difference between the between-level and within-level effects. Results suggest that BFLPEs existed in most of those 49 countries but the phenomena were different in those countries. The findings also suggest that when perceived relative standing to classmates was added to the model, BFLPEs decreased. Compared with previous cross-national BFLPE studies, this study demonstrated larger BFLPEs in the academic subject of math in individual countries.

Keywords: big-fish-little-pond-effect; cross-national study; multilevel latent variable modelling; TIMSS

Educational researchers are probably already familiar with the big-fish-little-pond-effect (BFLPE). Based on the social comparison theory, BFLPE was proposed by Marsh to understand the relationships between academic self-concepts and academic achievement. It is hypothesised that academic self-concepts are influenced by the ability levels of other students in the immediate context in addition to one's own ability level (e.g. Huguet et al., 2009; Marsh, 1987a; Marsh, Chessor, Craven, & Roche, 1995; Marsh, Kong, & Hau, 2000; Marsh, Seaton, et al., 2008). Students compare their own academic ability with the academic abilities of other students in their reference group and make judgements on how well they themselves are doing. For an average-ability student in a high-ability class or school, his/her academic achievement is below average and he/she will judge his/her academic ability as lower than average and thus below-average academic self-concept. For an average-ability student in a low-ability class or school, his/her academic achievement is above average and he/she will judge his/her academic ability as higher than average and thus higher academic self-concept. This suggests that students in high-ability classes and schools may have lower academic self-concepts than their equally able

*Email: WangZe@missouri.edu

counterparts in mixed- or low-ability classes and schools. In BFLPE, the central constructs are academic self-concept and academic ability. However, unlike achievement motivation models (e.g. expectancy-value model, self-determination theory and attribution theory) where psychological processes are proposed to understand academic performances and behaviours (e.g. Nagengast et al., 2011; Wang, Osterlind & Bergin, 2012), BFLPE focuses on how some salient, normative rankings (i.e. academic ability) affect self-evaluations of related attributes.

Recent cross-cultural studies have generalised BFLPE in different countries, using large-scale international databases (Marsh & Hau, 2003; Marsh, Trautwein, Lüdtke, & Koller, 2008; Seaton, Marsh, & Craven, 2009). Multilevel modelling is used as statistical analyses in those studies. The BFLPE research suggests that this effect exists in different cultures, and in both developed and developing countries (e.g. Seaton et al., 2009). The common treatment of academic self-concept measures in those cross-cultural BFLPE studies is standardising (z-scoring) the composite variables. This treatment is consistent with the classical true score model (Lord & Novick, 1968) and the variance of measurement error is assumed to be equal at different levels of true scores. In addition, different items carry equal weights in determining the amount of academic self-concept students have. Notably, two recent studies (Marsh et al., 2009; Nagengast & Marsh, 2011) proposed and used models to minimise measurement errors by using latent factors, instead of observed composites to represent academic self-concepts, in addition to the application of multilevel analysis to minimise sampling errors.

As a rich international database, the Trends in International Mathematics and Science Study (TIMSS) 2007 is the fourth data collection cycle by the International Association for the Evaluation of Educational Achievement. Although TIMSS is not designed to measure academic self-concept, after comparing some items included in the student questionnaire with items from other academic self-concept measures and contrasting them with measures of related constructs, it is most appropriate to label those items as measuring academic self-concept, rather than other self-constructs. In this present study, I examined the BFLPE in 49 countries using data from TIMSS 2007. Cross-national BFLPE research has been conducted using data from the Programme for International Student Assessment (PISA) database (Marsh & Hau, 2003; Seaton et al., 2009), where academic self-concept items are based on the Self-Description Questionnaire (SDQ) for adolescents (Marsh, 1992).

Of the existing BFLPE studies using cross-national data, Marsh, Trautwein, et al. (2008) conducted analyses on a total sample of 26 countries and used an overall academic self-concept measure and an academic achievement measure. Marsh and Hau (2003) conducted analyses both on a total sample of 26 countries and for individual countries, and used an overall academic self-concept measure and an academic achievement measure. Seaton et al. (2009) conducted analyses both on a total sample of 41 countries and for individual countries, and used a math self-concept measure and a math ability measure. Nagengast and Marsh (2011) conducted analyses for a total international sample of 57 countries, a total UK sample, and each of the four UK counties, and used a science self-concept measure and a science achievement measure.

In this present study, I used the self-concept and achievement measures for the subject of math since it has been demonstrated that academic self-concept is subject specific (Marsh, 1990c). Analyses were conducted for individual countries separately. Results from this study were compared with those from Marsh and Hau (2003) and from Seaton et al. (2009) where results for individual countries were available.

Self-concept

Self-concept beliefs rely on social comparative information and reflected appraisal from significant others (Bong & Skaalvik, 2003). Shavelson, Hubner, and Stanton (1976) refer to self-concept as a person's perception of himself or herself, and claim that this construct is organised, multifaceted, hierarchical, stable, developmental, evaluative and differentiable. Self-concept is both theoretically and empirically different from self-efficacy (Bong & Skaalvik, 2003; Pajares & Miller, 1994). Its key antecedents include: (a) frames of reference, (b) causal attributions, (c) reflected appraisals from significant others, (d) mastery experiences and (e) psychological centrality (Bong & Skaalvik, 2003). Self-concept in the academic domain, or academic self-concept, refers to individuals' perceptions or beliefs in their ability to do well in the academic domain. Academic self-concept has been demonstrated to be domain or subject specific in educational settings (Marsh, 1986, 1990a, 1990c). Therefore, researchers interested in self-concepts in particular subjects should use self-concept scales accordingly. Students' math self-concept refers to the representation and evaluation of their abilities in the academic domain of mathematics. Of the instruments that measure self-concept, the SDQ series (Marsh, 1987b, 1990b; Marsh & O'Neill, 1984) is widely used and has been shown to have good psychometric properties (Byrne, 1996). In both SDQI and SDQII, 10 items measure math self-concept. Arens, Yeung, Craven and Hasselhorn (2011) further divided the 10 items into the self-perception of competence component and the affect component. For the self-perception of competence component of math self-concept, there are five items: (a) I am good at mathematics; (b) I get good marks in mathematics; (c) work in mathematics is easy for me; (d) I learn things quickly in mathematics; and (e) I am dumb in mathematics. TIMSS 2007 includes items that assess self-beliefs without providing a theoretical rationale for including them. Students were asked to rate on a four-point Likert scale (1 = agree a lot, 2 = agree a little, 3 = disagree a little and 4 = disagree a lot). Of these items, three are consistent with the ability/competence component of math self-concept: (a) I usually do well in mathematics; (b) mathematics is not one of my strengths; and (c) I learn things quickly in mathematics. These three items consist of the math self-concept measure in this study. The two positively worded items were reversed coded so that a higher value meant more math self-concept.

Big-fish-little-pond-effect (BFLPE)

The BFLPE was proposed to understand the relationships between academic self-concepts and academic achievement. It is hypothesised that academic self-concepts are influenced by the ability levels of other students in the immediate context in addition to one's own ability level (e.g. Huguet et al., 2009; Marsh, 1987a; Marsh et al., 1995; Marsh et al., 2000; Marsh, Seaton, et al., 2008). Students in high-ability classes and schools may have lower academic self-concepts than their equally able counterparts in mixed-ability classes and schools. Since the idea of BFLPE was proposed and tested by Marsh, there have been numerous studies extending it to different educational settings and contexts, examining the mechanism of this effect through mediation analysis, and including additional student and environmental variables (Espenshade, Hale, & Chung, 2005; Marsh et al., 2000; Thijs, Verkuyten, & Helmond, 2010; Zeidner & Schleyer, 1999). Due to the availability of international databases, cross-national BFLPE research has begun to gain attention (Marsh & Hau, 2003; Marsh, Trautwein, et al., 2008; Seaton et al., 2009).

To understand BFLPE, we have to know the social comparison processes that serve as the theoretical support of it. According to the social comparison theory, individuals evaluate themselves by comparing themselves to others (Festinger, 1954; Suls, Martin, & Wheeler, 2002). To evaluate one's own ability, an individual makes use of information about a proxy such as the proxy's relative standing on attributes. When no objective standards are available, those in a person's immediate social group serve as the comparison target (Rogers, Smith, & Coleman, 1978). In schooling settings, usually the relative standing of all students in the immediate context (e.g. classes), especially on attributes related to academic learning, is salient. The average performance of all students in the immediate context may serve as the common comparison target when students make self-evaluations related to their academic ability. Based on the social comparison theory, earlier studies have found that academic ability within the classroom contributes to academic self-concept (Hay, Ashman, & van Kraayenoord, 1997; Rogers et al., 1978). Students compare their academic position with their classmates when they form their academic self-concept. As a result, students from different classes may have different self-evaluations even when their academic abilities are the same.

It is therefore worthwhile to separate the within-class comparison from the between-class comparison when students make self-perceptions of their math abilities. For the within-class comparison, students with higher math abilities have higher math self-concept (Hay et al., 1997). For the between-class comparison, class average math abilities can be modelled as a predictor of the class average math self-concept. This between-class effect has been termed as the assimilation effect or glory effect; in that students in a higher achieving immediate context feel proud and have higher math self-concept as a group than those in low-achieving contexts.

BFLPE assumes that in terms of academic ability and academic self-concept, the within-class comparison is larger than the between-class comparison (i.e. assimilation effect). The average performance of all students in the immediate context serves as the comparison standard for the within-class comparison. A student of high ability, *relative to this comparison standard*, would have a positive self-perception. The assimilation effect is the result that students compare their immediate context with other contexts. Students in a context with higher average ability would have a more positive average self-perception than students in a context with lower average ability. Both the within-class comparison and the between-class comparison effects would be positive. BFLPE is the net effect of the two (i.e. between-class comparison minus the within-class comparison) and is hypothesised to be negative (Marsh et al., 2009). The net effect, in multilevel modelling, is a contextual effect (i.e. the effect of the context after adjusting for the effect of individual characteristics, see Raudenbush & Bryk, 2002, p. 139).

Most BFLPE studies have focused on the absolute relative standing (i.e. students are compared to their classmates on test scores and grades), Huguet et al. (2009) found that students' *perceived* standing relative to most of their classmates and how students compare with their classmates are especially important in understanding the

BFLPE. That is, instead of using the static class average math ability for all students in the same class, how students *perceive* their academic standing, relative to their classmates is what matters to their academic self-concept. After controlling for perceived relative standing in class, the BFLPE disappeared (Huguet et al., 2009). One item from TIMSS 2007: *mathematics is more difficult for me than for many of my classmates*, was used as the perceived relative standing measure. Interestingly, this item, together with the three math self-concept items, are under one latent factor labelled 'Self-Confidence in Learning Mathematics' in TIMSS 2007 technical report (Martin & Preuschoff, 2008). While this item reflects the normative nature of self-concept beliefs, it is more closely related to the construct of 'Perceived Relative Standing' that began to appear in more recent BFLPE research. It is also demonstrated in the TIMSS 2007 technical report (Martin & Preuschoff, 2008) that the factor loading for this item is the lowest (.477) when it is modelled as an indicator of 'Self-Confidence in Learning Mathematics', together with the other three items (the loadings for the other three items from TIMSS technical report are .765, .653 and .812, respectively). TIMSS 2007 does not provide a theoretical rationale for including the items assessing self-beliefs.

It is worth noting that earlier BFLPE studies did not separate the within-class comparison from the between-class comparison. Instead, the between-class effect was the adjusted effect after controlling for the within-class effect; and therefore, the between-class effect in these studies *was* the BFLPE. In this study, the between-class effect was not after adjusting for the within-class effect. The difference in the between-class effect and within-class effect is the BFLPE. The statistical representation is in the 'Statistical Models' section below. It is also worth noting that the BFLPE is very specific to academic self-concept and largely non-significant for non-academic self-concept and general esteem (Marsh et al., 1995; Marsh & Hau, 2003).

TIMSS 2007

The complex data structure in TIMSS 2007 complicates statistical analysis. When TIMSS 2007 data are analysed, researchers have to consider two complex sample features: unequal selection probabilities (such as when selecting schools within each country) and clustering (such as when students were from the same class or school) (Muthen & Satorra, 1995; Stapleton, 2002, 2006). Clustering violates the independent observation assumption in traditional analysis techniques, such as in single-level confirmatory factor analysis (CFA). Standard errors of parameter estimates are biassed when clustering is ignored. Unequal selection probabilities will result in bias in the parameter estimates if ignored (Kaplan & Ferguson, 1999). In this study, weights at two levels (within classes and between classes) were specified to reflect TIMSS data collection design (for discussion of using sampling weights in multilevel modelling, see Asparouhov, 2004, 2006; Asparouhov & Muthen, 2006).

A matrix-sampling technique was used in the data collection stage for TIMSS 2007, with each student only responding to selected achievement assessment items. After data collection, students' math proficiency was estimated through item response theory, together with a multiple imputation technique (this procedure has been used for other large-scale projects such as the National Assessment of Educational Progress; for a more technical description of the procedure, see von Davier & Sinharay, 2007). By matrix sampling and multiple imputation, each student's math proficiency was represented with five plausible values, enough to allow for imputation error (Mislevy, 1991; Schafer, 1997) (however, also see Graham, Olchowski, & Gilreath, 2007 for more recent discussion on the number of imputations needed). The plausible values are not appropriate for use as individual student scores for reporting to the students; however, they can be used to estimate population characteristics, with better performance than point estimates (Wu, 2005). Our data analysis follows usual procedures for analysing multiple imputed data-sets, with each set of plausible values analysed separately and results from the five sets combined to allow for multiple imputation errors (Enders, 2010; Rubin, 1987).

The present study

In the present study, I tested the BFLPE on student math self-concept in 49 countries and compared the findings with earlier cross-national studies. A multilevel (i.e. two-level) latent variable modelling approach was applied to data in each country. The between-level units were classes and the within-level units were students. Three statistical models were analysed with data from each country. The first model was a two-level CFA model for math self-concept. The purpose of this model was to establish measurement invariance at the two levels. In addition, the variances of math self-concept at the within level and at the between level, as well as the intraclass correlation coefficient (ICC) of math self-concept, were obtained. It is hypothesised that multilevel measurement invariance exists for the math self-concept measure in each country. This model serves as the base model.

In the second statistical model, the BFLPE on math self-concept was tested for each of the 49 countries. At the within level, student math ability, relative to the class average, predicted the within-level math self-concept; a quadratic term of math ability was also included as a predictor of math self-concept, to be consistent with other cross-national studies on BFLPE. At the between level, class average math ability predicted the between-level math self-concept. The difference between the prediction at the between level and the prediction at the within level is the BFLPE. It is hypothesised that at both levels, math ability is positively related to math self-concept. Further, it is hypothesised that the between-level coefficient is smaller than the within-level coefficient (i.e. there is a BFLPE).

In the third model, relative standing perceived by student was added as another predictor of math self-concept at the within level. It is hypothesised that perceived relative standing is a positive predictor of math self-concept and that BFLPE in this model would be smaller than that in the second model, if not disappear completely.

Statistical analyses were conducted for the 49 countries separately. For cross-national comparisons, standardised effect sizes were calculated. Different effect sizes are anticipated due to differences in education systems and cultural and societal influences. However, no hypothesis is made regarding the existence or magnitude of the BFLPE in different countries and/or education systems. Thus, the present study is largely exploratory.

This is the first study that uses TIMSS 2007 data to examine the BFLPE in 49 countries and to include a perceived relative standing measure as well (Chiu, 2012 used TIMSS 2003 data to examine the BFLPE and the internal/external frame of reference model in 27 countries). In addition, this study uses a multilevel latent variable modelling approach to reduce both measurement error and sampling error (details are included in the Statistical Models section), and to separate within-class comparisons and between-class comparisons in each country.

Methods

Data and sample

This study used data from the 49 participating countries at the eighth-grade level of TIMSS 2007. The total sample consisted of 221,224 students from 10,275 classes and 7351 schools. More information about the sample is included in the TIMSS 2007 user guide (Foy & Olson, 2009).

As mentioned earlier, math self-concept is 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. Those three items are consistent with the definition of academic self-concept in that it indicates one's self-perceived *ability* within a given academic area (Bong & Skaalvik, 2003). Perceived relative standing is measured by one item: *Mathematics is more difficult for me than for many of my classmates*. The four items were rated on a 1–4 Likert-scale (1 = agree a lot, 2 = agree a little, 3 = disagree a little and 4 = disagree a lot) and positive statements were reverse coded. A higher value on the items means more math self-concept or more positively perceived relative standing. Each student had five plausible values of math proficiency. Weight variables consistent with TIMSS data collection design were used.

Analyses were conducted for each country separately. 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, schools and classes were specified as clustering variables and two-level analyses were performed.

Statistical models

BFLPE by definition is a contextual effect. Past research has used multilevel modelling formulation as the statistical evidence of this effect. Using the same notations as in Raudenbush and Bryk (2002), the statistical model of BFLPE can be written as the following, sometimes with variations.

$$\text{Level} - 1\,(\text{student level})\text{:sc}_{ij} = \beta_{0j} + \beta_{1j}(\text{ability}_{ij}) + r_{ij} \tag{1}$$

$$\text{Level} - 2\,(\text{class or school level})\text{:}\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{ability}_{.j}) + u_{0j} \tag{2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{ability}_{\text{j}}) + u_{1\text{j}} \tag{3}$$

where sc_{ij} and ability_{ij} are the observed academic self-concept and academic ability for student i at class/school j, respectively, often after z-scoring or *grand-mean centring*; and ability_j is the observed class/school average academic ability. With some joint distributional assumptions for the random components (r_{ij}, u_{0j} and u_{1j}), the parameter γ_{01}, together with other fixed effects, can be estimated and its statistical significance tested. The statistical evidence of BFLPE is that γ_{01} is statistically significantly less than zero. If students' academic ability is *group centred*, BFLPE exists when the *difference* between γ_{01} and γ_{10} is statistically significant. Raudenbush and Bryk (2002) also call this a compositional effect (p. 140). This way, BFLPE is decomposed into two components. The first component, the within-class component, is due to the fact that students within the same class are different. It reflects the *forced* relative standing of a student in the classroom. The second component, the between-class component, is due to the fact that students are in different classes.

Marsh et al. (2009) call this type of model doubly-manifest models in that the variables academic self-concept and academic ability contain measurement errors (they are scale scores or single indicator variables vs. latent factors) and that the aggregation of the level-1 variable academic ability at level 2 is manifest (i.e. there is sampling error contained in the aggregation). Variations of the above model include adding a quadratic term of academic ability, adding a third level (e.g. country), and controlling for other student characteristics and/or class/school characteristic variables.

The multilevel latent covariate model (Lüdtke et al., 2008) is a more sophisticated treatment of contextual effects. In this approach, both the predictor and the outcome variable are decomposed into unobserved components. If both variables are measured at Level 1, the decomposition is:

$$X_{ij} = \mu_x + U_{xj} + R_{xij} \tag{4}$$

$$Y_{ij} = \mu_y + U_{yj} + R_{yij} \tag{5}$$

where μ_x is the grand mean of X and is a constant. U_{xj} are group-specific deviations and R_{xij} are individual deviations within group. Similar notations are for variable Y. X_{ij} and Y_{ij} are observed variables, and U_{xj}, R_{xij}, U_{yj} and R_{yij} are unobserved variables. The relationships between X and Y are specified at the within and between levels via relationships between the unobserved terms. That is:

$$R_{yij} = \beta_{\text{within}} R_{xij} + \varepsilon_{ij} \tag{6}$$

$$U_{yj} = \beta_{\text{between}} U_{xj} + \delta_j \tag{7}$$

In Marsh et al. (2009), the latent covariate model was further extended to incorporate predictors and outcome variables as latent factors and the term *doubly-latent models* is used to refer to models that both use the multilevel latent covariate approach and minimise measurement errors through latent aggregation. Measurement errors in scale scores can be minimised by including latent factors instead of observed composite scores. *Implicit group mean centring* can help reduce sampling errors and is readily available in Mplus. In this study, math self-concept was a latent factor with three indicators. With the implicit group mean centring, math self-concept is decomposed:

$$\text{sc}_{ij} = \text{sc}_{wij} + \text{sc}_{bj} \tag{8}$$

where sc_{ij} is the latent math self-concept for student i in group j; sc_{wij} is its within component and sc_{bj} is its between component, corresponding to R_{yij} and U_{yj} in Equations (4) and (5), respectively. The grand mean of sc_{ij} is arbitrary and set to zero. sc_{wij} and sc_{bj} are assumed to be measured by three indicators with the following equations:

$$\begin{aligned} Y_{1ij} &= \alpha_{1j} + \lambda_1 \text{sc}_{wij} + e_{1ij} \\ Y_{2ij} &= \alpha_{2j} + \lambda_2 \text{sc}_{wij} + e_{2ij} \\ Y_{3ij} &= \alpha_{3j} + \lambda_3 \text{sc}_{wij} + e_{3ij} \end{aligned}$$

$$\alpha_{1j} = \lambda_1 \mathrm{sc}_{bj} + r_{1j}$$
$$\alpha_{2j} = \lambda_2 \mathrm{sc}_{bj} + r_{2j}$$
$$\alpha_{3j} = \lambda_3 \mathrm{sc}_{bj} + r_{3j}$$

α_{1j}, α_{2j} and α_{3j} are random variables representing group-specific indicator intercepts at the within level and serve as indicators of the latent factor sc_{bj} at the between level. The loadings corresponding to each observed variable Y are invariant across levels. This ensures that the interpretation of math self-concept at the within and between levels are similar. Combining the above equations, we have

$$Y_{1ij} = \lambda_1(\mathrm{sc}_{wij} + \mathrm{sc}_{bj}) + r_{1j} + e_{1ij} = \lambda_1 \mathrm{sc}_{ij} + r_{1j} + e_{1ij}$$
$$Y_{2ij} = \lambda_2(\mathrm{sc}_{wij} + \mathrm{sc}_{bj}) + r_{2j} + e_{2ij} = \lambda_2 \mathrm{sc}_{ij} + r_{2j} + e_{2ij}$$
$$Y_{3ij} = \lambda_3(\mathrm{sc}_{wij} + \mathrm{sc}_{bj}) + r_{3j} + e_{3ij} = \lambda_3 \mathrm{sc}_{ij} + r_{3j} + e_{3ij}$$

with $\mathrm{math}_{ij} = \mu_{\mathrm{math}} + \mathrm{math}_{wij} + \mathrm{math}_{bij}$, we are interested in the following relationships:

$$\mathrm{sc}_{wij} = \beta_{\mathrm{within}} \mathrm{math}_{wij} + \beta_2 (\mathrm{math}_{ij})^2 + \varepsilon_{ij} \tag{9}$$

$$\mathrm{sc}_{bj} = \beta_{\mathrm{between}} \mathrm{math}_{bj} + \delta_j \tag{10}$$

In fact, each student in TIMSS 2007 has five plausible values of math ability; the model can be performed five times to obtain five sets of estimates. Therefore, instead of using latent factor, the measurement error of math ability is accounted for by using plausible values.

In order to compare the contextual effects across countries, some dimensionless effect size has to be used. In multilevel modelling, one commonly used effect size for continuous predictors is described by Tymms (2004). In this study, the effect size of math ability on math self-concept at the within and between levels are described in the following equations, respectively:

$$\mathrm{ES}_{\mathrm{within}} = 2 \times \beta_{\mathrm{within}} \times \mathrm{SD}_{\mathrm{math}_w} / \mathrm{SD}_{\mathrm{sc}_{ij}} \tag{11}$$

$$\mathrm{ES}_{\mathrm{between}} = 2 \times \beta_{\mathrm{between}} \times \mathrm{SD}_{\mathrm{math}_b} / \mathrm{SD}_{\mathrm{sc}_{ij}} \tag{12}$$

where $\mathrm{SD}_{\mathrm{math}_w}$ and $\mathrm{SD}_{\mathrm{math}_b}$ are the standard deviation of the within and between math ability, and $\mathrm{SD}_{\mathrm{sc}_{ij}}$ is the standard deviation of the total math self-concept calculated as

$$\mathrm{SD}_{\mathrm{sc}_{ij}} = \sqrt{\beta^2_{\mathrm{within}} \times \mathrm{Var}_{\mathrm{math}_w} + \beta^2_2 \times \mathrm{Var}_{(\mathrm{math}_{ij})^2} + \mathrm{Var}_{\varepsilon_{ij}} + \beta^2_{\mathrm{between}} \times \mathrm{Var}_{\mathrm{math}_b} + \mathrm{Var}_{\delta_j}} \tag{13}$$

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

$$\mathrm{ES}_{\mathrm{BFLPE}} = 2 \times (\beta_{\mathrm{between}} - \beta_{\mathrm{within}}) \times \mathrm{SD}_{\mathrm{math}_b} / \mathrm{SD}_{\mathrm{sc}_{ij}} \tag{14}$$

When the additional predictor of *perceived* relative standing is added to the within-level model in (9), the calculation is modified accordingly and includes the contribution of the additional variable.

In this study, there were 49 (countries) × 5 (plausible values for math ability) = 245 data-sets. Three different models were conducted with each of those 245 data-sets. Results from the 5 data-sets representing the same country were later combined using procedures described in Rubin (1987) and Enders (2010). Due to the number of data-sets, running each model separately with each data-set seems tedious and is prone to error. An R package, 'MplusAutomation' written by Hallquist (2011), is used to automate the analysis. The following steps describe the data and model process.

Step 1. Process TIMSS 2007 data downloaded to generate the 245 data-sets with identical format and variable names.

Step 1a. The data downloaded are student background data files with identification of country, school, class and student. The five plausible values of student math ability were also included. The data are SPSS data-sets with a name 'bsg***m4.sav', where '***' are the three-letter code for each of the 49 countries. Data process in SPSS included renaming identification variables, reverse coding, replacing missing values with 9s for the math self-concept items, and formatting variables. The data file for each country was output as a '***.dat' where '***' is the three-letter country code. In this .dat file, country, school, class and student ids were kept, together with math self-concept items, math ability plausible values, and weight variables and adjusting factors.

Step 1b. A SAS macro was written to process those .dat files from Step 1a. This step included generating five data-sets from each of the .dat file and standardising variables. The five data-sets from the same .dat had the same variables and values except for the math ability variable which represented a different series in each of the five data-sets. During this step, care was taken so that values of 9 on math self-concept items were set to missing before standardisation and those missing values were set to be 9s in the final output data files. The SAS macro was called 49 times for the 49 countries. This step generated 245 datafiles with names 'z***math#.dat', where '***' is the three-letter country code and # has values 1, 2, 3, 4 or 5, indexing the five series of plausible values.

Step 2. Models were run using the R package 'MplusAutomation', which calls Mplus to run models and model results were processed in R.

Step 2a. A Mplus template was created for each statistical model. In the template file, the 245 data files were specified as iterations so that 245 Mplus syntax files were created using the 'createModels' function in the R package 'MplusAutomation'. The 245 generated Mplus syntax files were identical except for their names and the data-set specified.

Step 2b. The 'runModels' function in the R package 'MplusAutomation' was used to run the 245 syntax files, producing 245 Mplus output files. Parameter estimates were extracted from those 245 Mplus output files.

Step 2c. Those parameter estimates from Step 2b were further processed to obtain results for each country, averaging over the five plausible values. The standard error of each interested parameter estimate was calculated from both the within-plausible values and between-plausible values.

During the data process and data analysis steps, country codes and numbering of plausible values were carried when results are further processed and they were checked to make sure that the orders were consistent.

Statistical analysis

All variables were standardised for the separate analyses. Standardisation of plausible values of math achievement was within each series. Sampling weights were applied at both the between and the within levels. The between-level weight is the reciprocal of the probability that a class is sampled, adjusting for non-participating schools and classrooms. The within-level weight is the reciprocal of the probability that a student is sampled, *given the class is sampled*, adjusting for non-participating students.

For each country, three models were tested: the first model only included the three math self-concept items and tested the measurement invariance model fit across levels. In addition, the ICC of the latent factor was obtained. This model serves as the baseline model. The second model included the linear and quadratic effects of student-level achievement and the linear effect of the class-level achievement. All models were specified using implicit group-mean centring. BFLPE as a contextual effect presents if the between-level regression coefficient is significantly different from the within-level regression coefficient (Enders & Tofighi, 2007; Marsh et al., 2009). The difference between the two coefficients was specified as an additional parameter and hypothesis testing of it was testing the presence of BFLPE. In the third model, perceived relative standing was added as an additional predictor of math self-concept at the within level. The implicit group-mean centring was used. This model tests whether BFLPE was reduced or even offset after controlling for students' perceived relative standing among classmates.

Results

Multilevel CFA of math self-concept (Model 1)

Following the procedure of testing measurement invariance at the within and between levels (Nagengast & Marsh, 2011), the multilevel CFA model was fitted to data in each of the 49 countries separately. It should be noted that the multilevel CFA model assuming only configural invariance is saturated with only three indicators. The model fit indices assuming factor loading invariance (Model 1) is in Table 1. The model fit in Botswana suggested no multilevel measurement invariance. Interestingly, in the USA, despite high CFI and TLI, the chi-square statistic was much larger than that for other countries and the RMSEA only indicated marginally acceptable model fit. Table 1 also includes the within and between variance of math self-concept and the ICC. There are seven countries with statistically non-significant ($p > .01$) between-level math self-concept variance. The smallest ICC was in Armenia (1.8%) and the largest was in Indonesia (22.8%).

BFLPE (Model 2)

The second model included the linear and quadratic effects of student-level achievement and the linear effect of the class-level achievement. All models were specified using implicit group-mean centring. BFLPE as a contextual effect presents

Table 1. Model fit statistics for multilevel CFA with multilevel measurement invariance of math self-concept in 49 countries.

Code	Country	# Classes	# Students	Chi-square	CFI	TLI	RMSEA	Variance of within-level SC	Variance of between-level SC	ICC
arm	Armenia	250	4689	.56	1.00	1.00	.00	.544	.010[a]	.0181
aus	Australia	238	4069	5.09	1.00	1.00	.02	.581	.095	.1405
bgr	Bulgaria	247	4019	17.05	.98	.95	.04	.476	.070	.1282
bhr	Bahrain	201	4230	.29	1.00	1.01	.00	.123	.006	.0465
bih	Bosnia and Herzegovina	181	4220	2.10	1.00	1.00	.00	.553	.037	.0627
bwa	Botswana	151	4208	50.53	.93	.78	.08	.241	.017[a]	.0659
col	Colombia	149	4873	2.21	1.00	1.00	.00	.594	.077	.1148
cyp	Cyprus	259	4399	1.38	1.00	1.00	.00	.601	.026	.0415
cze	Czech Republic	212	4845	4.19	1.00	1.00	.01	.699	.028	.0385
dza	Algeria	149	5447	4.07	1.00	.99	.01	.085	.006	.0659
egy	Egypt	238	6582	8.20	.96	.89	.02	.122	.010[a]	.0758
eng	England	238	4025	78.51	.96	.88	.10	.506	.117	.1878
geo	Georgia	184	4178	.51	1.00	1.01	.00	.429	.039	.0833
gha	Ghna	174	5294	.32	1.00	1.01	.00	.324	.051	.1360
hkg	Hong Kong SAR	120	3470	5.98	1.00	.99	.02	.577	.029	.0479
hun	Hungary	246	4111	11.84	.99	.98	.04	.641	.048	.0697
idn	Indonesia	149	4203	.22	1.00	1.01	.00	.319	.094[a]	.2276
irn	Iran, Islamic Rep. of	208	3981	1.35	1.00	1.00	.00	.463	.067	.1264
isr	Israel	146	3294	4.64	1.00	.99	.02	.553	.024	.0416
ita	Italy	287	4408	4.48	1.00	1.00	.02	.644	.039	.0571
jor	Jordan	200	5251	2.31	1.00	1.00	.01	.247	.034	.1210
jpn	Japan	169	4312	3.54	1.00	1.00	.01	.622	.026	.0401
kor	Korea, Rep. of	150	4240	.27	1.00	1.00	.00	.791	.021	.0259
kwt	Kuwait	158	4091	18.88	.97	.92	.05	.055	.002[a]	.0351
lbn	Lebanon	205	3786	10.71	.98	.95	.03	.474	.057	.1073
ltu	Lithuania	258	3991	13.50	1.00	.99	.04	.604	.039	.0607
mar	Morocco	131	3060	.51	1.00	1.01	.00	.086	.010	.1042
mlt	Malta	232	4670	.72	1.00	1.00	.00	.615	.141	.1865

(*Continued*)

Table 1. (*Continued*).

Code	Country	# Classes	# Students	Chi-square	CFI	TLI	RMSEA	Variance of within-level SC	Variance of between-level SC	ICC
mng	Mongolia	152	4499	12.46	.99	.97	.04	.49	.056	.1026
mys	Malaysia	163	4466	15.40	.98	.95	.04	.412	.062	.1308
nor	Norway	264	4627	5.03	1.00	1.00	.02	.626	.016	.0249
omn	Oman	158	4752	1.09	1.00	1.01	.00	.143	.007[a]	.0467
pse	Palestinian Nat'l Auth.	153	4378	.57	1.00	1.01	.00	.27	.024	.0816
qat	Qatar	288	7184	3.75	1.00	.99	.01	.10	.007	.0654
rom	Romania	266	4198	19.16	.98	.94	.05	.539	.056	.0941
rus	Russian Federation	271	4472	1.15	1.00	1.00	.00	.568	.054	.0868
sau	Saudi Arabia	204	4243	3.17	1.00	.99	.01	.102	.006[a]	.0556
scg	Serbia	227	4045	7.45	1.00	.99	.03	.694	.045	.0609
sco	Scotland	244	4070	25.72	.98	.95	.05	.492	.095	.1618
sgp	Singapore	326	4599	5.93	1.00	1.00	.02	.688	.117	.1453
svn	Slovenia	260	4043	1.34	1.00	1.00	.00	.57	.032	.0532
swe	Sweden	307	5215	16.90	1.00	.99	.04	.597	.024	.0386
syr	Syrian Arab Republic	150	4650	.26	1.00	1.01	.00	.128	.011	.0791
tha	Thailand	150	5412	2.49	1.00	1.00	.01	.474	.052	.0989
tun	Tunisia	169	4080	35.03	.97	.93	.06	.298	.028	.0859
tur	Turkey	146	4498	3.09	1.00	1.00	.01	.577	.046	.0738
twn	Chinese Taipei	153	4046	4.62	1.00	1.00	.02	.629	.04	.0598
ukr	Ukraine	184	4424	32.29	.98	.95	.06	.497	.062	.1109
usa	USA	510	7377	124.25	.98	.94	.09	.583	.103	.1501

Note: In the model for each country, there are 13 free parameters and 2 degrees of freedom. SC = math self-concept. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; ICC = Intraclass correlation coefficient.
[a] $p > .01$.

Table 2. Effect sizes in 49 countries.

	Model 2			Model 3					
Country	Within-level linear effect of math ability	Between-level effect of math ability	Contextual effect (BFLPE)	Within-level linear effect of math ability	Between-level effect of math ability	Effect of perceived relative standing	Contextual effect (BFLPE)	BFLPE in Marsh & Hau (2003)[a]	BFLPE in Seaton, et al. (2009)
Armenia	$.46^{***}$	.08	$-.17^{**}$	$.36^{***}$	.07	$.83^{***}$	$-.13^{*}$	–	–
Australia	1.13^{***}	$.59^{***}$	$-.69^{***}$	$.84^{***}$	$.53^{***}$	$.96^{***}$	$-.42^{***}$	–	$-.281^{*}$
Bulgaria	$.78^{***}$	$.41^{***}$	$-.31^{**}$	$.62^{***}$	$.3^{**}$	$.67^{***}$	$-.27^{*}$	–	–
Bahrain	1.22^{***}	.14	$-.48^{***}$	$.89^{***}$	−.01	1.16^{***}	$-.46^{***}$	–	–
Bosnia and Herzegovina	1.08^{***}	$.17^{**}$	$-.34^{***}$	$.74^{***}$	.11	$.97^{***}$	$-.25^{***}$	–	–
Botswana	$.49^{***}$	$-.24^{***}$	$-.49^{***}$	$.37^{***}$	$-.28^{***}$	$.47^{***}$	$-.47^{***}$	–	–
Colombia	$.79^{***}$	.05	$-.62^{***}$	$.65^{***}$	−.01	$.67^{***}$	$-.56^{***}$	–	–
Cyprus	1.12^{***}	.08	$-.15^{*}$	$.82^{***}$	.07	$.92^{***}$	$-.11^{*}$	–	–
Czech Republic	1.28^{***}	$.07^{**}$	$-.82^{***}$	$.95^{***}$	.02	1.01^{***}	$-.63^{***}$	$-.24^{*}$	$-.446^{*}$
Algeria	1.02^{***}	$.26^{**}$	−.06	$.93^{***}$	$.20^{*}$	$.71^{***}$	−.09	–	–
Egypt	$.95^{***}$	$.34^{***}$	$-.25^{**}$	$.79^{***}$	$.28^{**}$	$.77^{***}$	$-.20^{**}$	–	–
England	$.81^{***}$	$.72^{***}$	$-.79^{***}$	$.51^{***}$	$.57^{***}$	1.06^{***}	$-.37^{***}$	$-.23^{*}$ (UK)	$-.344^{*}$ (UK)
Georgia	$.95^{***}$	.09	$-.54^{***}$	$.79^{***}$	.11	$.74^{***}$	$-.41^{***}$	–	–
Ghana	$.61^{***}$	.03	$-.47^{***}$	$.56^{***}$	−.02	$.4^{***}$	$-.49^{***}$	–	–
Hong Kong SAR	$.88^{***}$	$.45^{***}$	$-.75^{***}$	$.6^{***}$	$.30^{***}$	1.10^{***}	$-.52^{***}$	–	$-.200^{*}$
Hungary	1.28^{***}	$.26^{***}$	$-.69^{***}$	$.92^{***}$	$.20^{**}$	$.99^{***}$	$-.49^{***}$	−.05	$-.323^{*}$
Indonesia	$.13^{*}$	$-.69^{***}$	$-.81^{***}$	.05	$-.68^{***}$	$.36^{***}$	$-.73^{***}$	–	$-.235^{*}$
Iran, Islamic Rep. of	1.01^{***}	$.26^{***}$	$-.51^{***}$	$.76^{***}$	.18	$.95^{***}$	$-.4^{***}$	–	–
Israel	$.92^{***}$	.13	$-.52^{***}$	$.69^{***}$	−.03	$.95^{***}$	$-.52^{***}$	–	–
Italy	1.20^{***}	−.01	$-.78^{***}$	$.88^{***}$	−.03	1.02^{***}	$-.59^{***}$	$-.36^{*}$	$-.409^{*}$
Jordan	1.10^{***}	$.48^{***}$	$-.26^{**}$	$.93^{***}$	$.38^{**}$	$.84^{***}$	$-.25^{***}$	–	–

(*Continued*)

Table 2. (*Continued*).

	Model 2			Model 3					
Country	Within-level linear effect of math ability	Between-level effect of math ability	Contextual effect (BFLPE)	Within-level linear effect of math ability	Between-level effect of math ability	Effect of perceived relative standing	Contextual effect (BFLPE)	BFLPE in Marsh & Hau (2003)[a]	BFLPE in Seaton, et al. (2009)
Japan	1.20^{***}	.12	$-.45^{***}$	1.01^{***}	$.12^{*}$	$.84^{***}$	$-.36^{***}$	–	$-.307^{*}$
Korea, Rep. of	1.35^{***}	$.22^{***}$	$-.22^{***}$	$.90^{***}$	$.19^{***}$	1.20^{***}	$-.10^{***}$	−.02	−.014
Kuwait	1.00^{***}	$.13^{*}$	$-.35^{***}$	$.65^{***}$	.04	1.21^{***}	$-.27^{***}$	–	–
Lebanon	$.91^{***}$	$.36^{***}$	$-.53^{***}$	$.80^{***}$	$.26^{*}$	$.55^{***}$	$-.53^{***}$	–	–
Lithuania	1.25^{***}	$.29^{***}$	$-.48^{***}$	$.95^{***}$	$.26^{**}$	$.89^{***}$	$-.33^{***}$	–	–
Morocco	$.94^{***}$	$.31^{**}$	−.22	$.58^{***}$	.19	1.26^{***}	−.13	–	–
Malta	1.02^{***}	$.58^{***}$	$-.91^{***}$	$.82^{***}$	$.46^{***}$	$.78^{***}$	$-.75^{***}$	–	–
Mongolia	$.63^{***}$	.00	$-.42^{***}$	$.54^{***}$	$-.13^{*}$	$.75^{***}$	$-.49^{***}$	–	–
Malaysia	$.79^{***}$	$.56^{***}$	$-.5^{***}$	$.68^{***}$	$.48^{***}$	$.69^{***}$	$-.44^{***}$	–	–
Norway	1.30^{***}	$.18^{***}$	$-.28^{***}$	$.96^{***}$	.13	$.97^{***}$	$-.21^{***}$	$-.18^{*}$	$-.168^{*}$
Oman	1.34^{***}	$.33^{***}$	$-.41^{***}$	1.12^{***}	$.26^{**}$	$.92^{***}$	$-.36^{***}$	–	–
Palestinian Nat'l Auth.	1.20^{***}	$.19^{**}$	$-.44^{***}$	1.18^{***}	$.17^{*}$	$.35^{***}$	$-.44^{***}$	–	–
Qatar	$.98^{***}$	.19	$-.45^{***}$	$.70^{***}$	.10	1.16^{***}	$-.37^{***}$	–	–
Romania	$.99^{***}$	$.26^{**}$	$-.47^{***}$	$.82^{***}$	.20	$.58^{***}$	$-.4^{***}$	–	–
Russian Federation	1.17^{***}	$.31^{***}$	$-.64^{***}$	$.71^{***}$	.17	1.27^{***}	$-.41^{***}$	$-.21^{*}$	$-.187^{*}$
Saudi Arabia	1.10^{***}	$.19^{*}$	$-.32^{**}$	1.01^{***}	.16	$.54^{***}$	$-.31^{*}$	–	–
Serbia	1.34^{***}	$.29^{***}$	$-.37^{***}$	1.07^{***}	.21	$.85^{***}$	$-.32^{***}$	–	$-.181^{*}$
Scotland	$.77^{***}$	$.68^{***}$	$-.42^{***}$	$.55^{***}$	$.57^{***}$	1.02^{***}	$-.21^{*}$	–	–
Singapore	$.94^{***}$	$.59^{***}$	-1.14^{***}	$.71^{***}$	$.45^{***}$	$.95^{***}$	$-.85^{***}$	–	–
Slovenia	1.13^{***}	$.19^{*}$	$-.20^{*}$	$.86^{***}$	.13	$.84^{***}$	$-.18^{**}$	–	–
Sweden	1.22^{***}	$.17^{***}$	$-.34^{***}$	$.77^{***}$	.09	1.19^{***}	$-.23^{***}$	$-.33^{*}$	$-.202^{*}$
Syrian Arab Republic	1.07^{***}	$.27^{***}$	$-.52^{***}$	$.89^{***}$	$.22^{*}$	1.04^{***}	$-.45^{***}$	–	–

(*Continued*)

Table 2. (*Continued*).

	Model 2			Model 3					
Country	Within-level linear effect of math ability	Between-level effect of math ability	Contextual effect (BFLPE)	Within-level linear effect of math ability	Between-level effect of math ability	Effect of perceived relative standing	Contextual effect (BFLPE)	BFLPE in Marsh & Hau (2003)[a]	BFLPE in Seaton, et al. (2009)
Thailand	$.50^{***}$	.08	$-.44^{***}$	$.43^{***}$	.05	$.48^{***}$	$-.4^{***}$	–	$-.194^{*}$
Tunisia	1.00^{***}	$.27^{***}$	$-.17^{**}$	$.72^{***}$	.15	1.1^{***}	$-.16^{***}$	–	$-.161^{*}$
Turkey	1.12^{***}	$.17^{**}$	$-.72^{***}$	$.91^{***}$	.11	$.89^{***}$	$-.61^{***}$	–	$-.252^{*}$
Chinese Taipei	1.22^{***}	$.45^{***}$	$-.22^{***}$	$.89^{***}$	$.32^{***}$	1.11^{***}	$-.17^{***}$	–	–
Ukraine	1.03^{***}	$.37^{***}$	$-.30^{***}$	$.69^{***}$	$.27^{***}$	1.02^{***}	$-.17^{**}$	–	–
USA	$.87^{***}$	$.57^{***}$	$-.48^{***}$	$.51^{***}$	$.40^{***}$	1.27^{***}	$-.21^{***}$	$-.26^{*}$	$-.230^{*}$

Note: [a]A total test score of verbal, math and science achievement was used.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

if the between-level regression coefficient is significantly different from the within-level regression coefficient (Enders & Tofighi, 2007; Marsh et al., 2009). The difference between the two coefficients was specified as an additional parameter and hypothesis testing of it was testing the presence of BFLPE.

The within-level effect of math ability on math self-concept (β_{within}) was positive and statistically significant (p = .017 in Indonesia, and p < .00001 in all other countries). At the .01 level, the between-level effect of math ability of math self-concept on math self-concept (β_{between}) was not statistically significant in 15 countries, was positive and statistically significant in 32 countries, and was negative and statistically significant in 2 countries (Botswana and Indonesia); the difference between β_{between} and β_{within} (i.e. BFLPE) was negative and statistically significant in all but four countries (Cyprus, Algeria, Morocco and Slovenia). The effect sizes of β_{within}, β_{between} and BFLPE (i.e. $\beta_{\text{between}} - \beta_{\text{within}}$) are shown in Table 2 as well as in Figure 1. In Figure 1, the countries are ordered by the effect size of the BFLPE.

BFLPE and effect of perceived relative standing on math self-concept (Model 3)

In the third model, perceived relative standing was added as an additional within-level predictor. That is, this model included the linear and quadratic effects of student-level achievement and the linear effect of perceived relative standing at the within level, and the linear effect of the school-level achievement at the between level. The implicit group-mean centring was used. This model tests whether BFLPE was reduced or even offset after controlling for students' perceived relative standing among classmates.

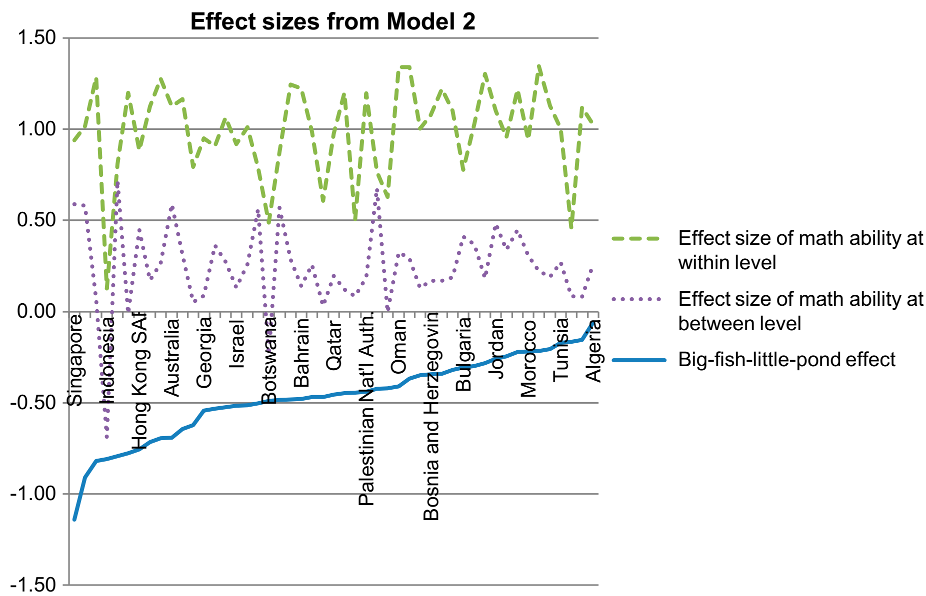

Figure 1. Effect sizes of math ability on math self-concept at the within level and at the between level, as well as effect size of the BFLPE, by country. The BFLPE is the contextual effect calculated as the difference between between-level math ability effect and within-level math ability effect.

Perceived relative standing was positively related to student math self-concept at the within level. 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 student compare himself/herself to the classmates. Effect sizes from this model are shown in Table 2 as well as in Figure 2. The order of countries is the same as that in Figure 1.

The effect of perceived standing relative to classmates on math self-concept was positive and statistically significant in all the 49 countries. Controlling for perceived relative standing, the within-level effect of math ability on math self-concept (β_{within}) was positive and statistically significant with *p*s < .001 in all countries but Indonesia (*p* = .277). At the .01 level, the between-level effect of math ability on math self-concept (β_{between}) was not statistically significant in 30 countries, was positive and statistically significant in 17 countries, and was negative and statistically significant in 2 countries (Botswana and Indonesia); the difference between β_{between}

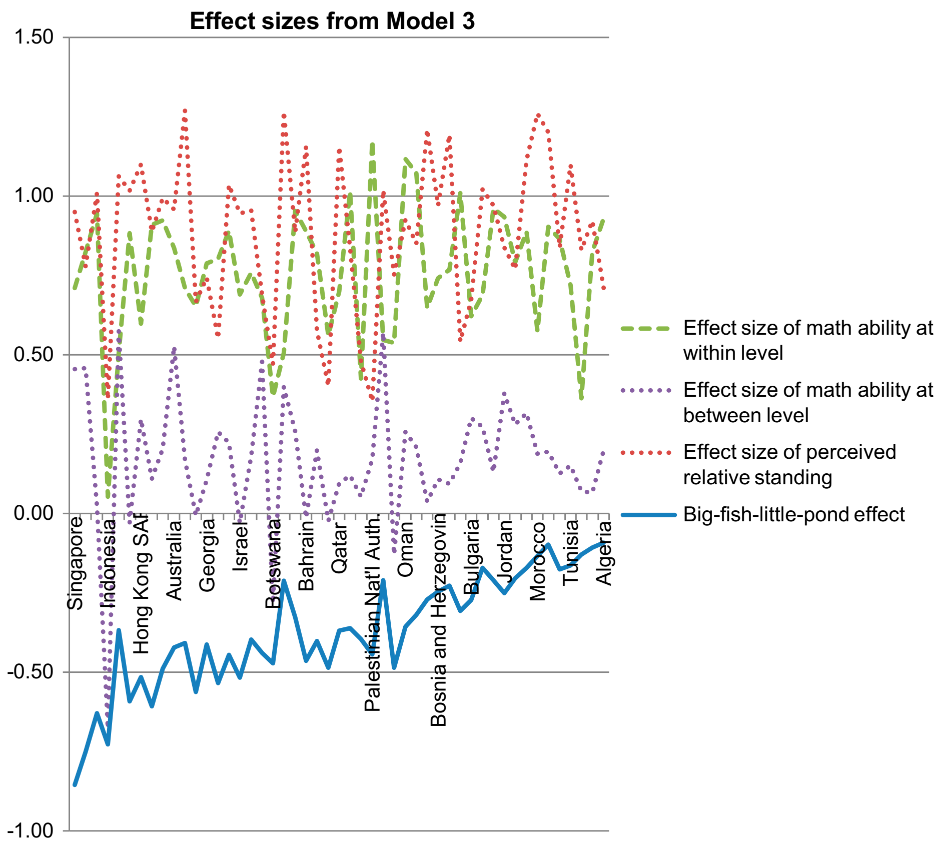

Figure 2. Effect sizes of math ability on math self-concept at the within level and at the between level, effect size of perceived relative standing on math self-concept, and effect size of the BFLPE, by country. The BFLPE is the contextual effect calculated as the difference between between-level math ability effect and within-level math ability effect, controlling for the perceived relative standing effect.

and β_{within} (i.e. that is, the BFLPE) controlling for perceived relative standing, was statistically significant in 42 countries.

From this model, perceived relative standing does not offset the BFLPE, although it does reduce BFLPE in most countries. While there are not many changes in term of significance levels, the effect sizes were decreased when perceived relative standing was added at the within level in all but three countries (in Algeria, BFLPE changed from .06 to .09 and remained non-significant; in Ghana, BFLPE changed from .47 to .49; and in Mongolia, BFLPE changed from .42 to .49). Interestingly, after controlling for perceived relative standing in class, the effect of math ability at the between level became non-significant or decreased (with three exceptions: in Botswana, the negative effect became larger; in Japan, the non-significant effect became significant at the .05 level without change in the effect size; in Mongolia, the non-significant effect became negative at the .05 level).

Discussion

Using doubly-latent models, this study examined the contextual effect of math ability on math self-concept, using a different database than the one used in previous cross-cultural BFLPE studies (Marsh & Hau, 2003; Marsh, Trautwein, et al., 2008; Seaton et al., 2009). This contextual effect, if negative, has been called BFLPE. In addition, another construct that has begun to gain attention in BFLPE research, *perceived relative standing*, was modelled to see if it would reduce or offset BFLPE.

The finding that BFLPE existed in most of the 49 countries examined is in general consistent with previous cross-national studies. However, this study is different from previous ones in several ways. First, it takes into account measurement error and sampling error together as recommended by Marsh et al. (2009). The measurement error in math self-concept was accounted for by a latent within- and a latent between-level factor. The measurement error in math ability was accounted for by using five series of plausible values. The sampling error was accounted for by taking into account the complex data structure together with an implicit-mean centring approach. Second, although BFLPE existed in multiple countries, the phenomenon was different. As a contextual effect, BFLPE statistically means that the effect at the between level is different from the pure within-level effect. This study modelled the within- and between-level effects directly. In all 49 countries, the pure within-level effect existed, consistent with earlier intraclass findings and the frame-of-reference notion in social comparison theory (Hay et al., 1997; Rogers et al., 1978). In contrast, the between-level effect did not exist in 12 countries at the .05 level, suggesting that a class with a higher average math ability does not necessary have a higher average math self-concept. The BFLPE, modelled as the between- and within-level difference, may be due to the existence and non-existence of effects at the two levels, or to the different magnitude of effects at the two levels. For example, in Armenia, the BFLPE was due to non-existence of between-level effect and positive within-level effect. In contrast, in Bulgaria, the BFLPE was due to smaller between-level effect than within-level effect. Surprisingly, in two countries (Botswana and Indonesia), the between-level effect was negative. Translating this statistic into plain English, this means that students in higher ability classes in general had lower math self-concept (note, this is *not* BFLPE which suggests that students of *similar ability* in a higher ability class has lower math self-concept than in a lower ability class). This negative between-level effect could also result from lack of multilevel measurement invariance and/or small between-level math self-concept variance in these two countries.

The largest BFLPE was in Singapore (−1.14) and the smallest was in Algeria (−0.06 with $p > .05$). In Singapore, being in a class with average math ability one standard deviation above the mean of average math abilities for all classes was associated with about 1.14 standard deviations below the mean of all students in all classes on math self-concept. In Algeria, being in a class with higher average math ability was not significantly associated with a decrease in math self-concept.

In Table 2, the BFLPEs reported in Marsh and Hau (2003) and Seaton et al. (2009) are also included. Nine countries overlapped between the present study and Marsh and Hau (2003). BFLPE was observed in all the nine countries in the present study. In Marsh and Hau (2003), it was not observed in Hungary or Korea. It should be noted that in Marsh and Hau (2003), academic self-concept was not differentiated by subject and academic achievement was a total score of verbal, math and science subjects; while in this present study, only math self-concept and math achievement was used. Seventeen countries overlapped between the present study and Seaton et al. (2009). BFLPEs in those 17 countries were observed from both studies except for Korea. In Korea, BFLPE was observed in this present study but was not statistically significant in Seaton et al. (2009). Seaton et al. (2009) argue that statistically non-significant BFLPE in Korea in their study may be due to reflected glory effect. In this study, since the BFLPE was the contextual effect modelled as the difference between the between-level effect of math ability and the within-level effect of math ability on math self-concept, it is possible to identify the cause of this difference. Of the 49 countries, Korea had the largest within-level effect of math ability (1.35) and relatively small between-level effect (.22), suggesting that in Korea, the pure effect of within-class relative standing was stronger than that of class average – that is, when students evaluate their math ability, they put more emphasis on how they compare to those in the immediate context (i.e., classes in this study) than on comparisons with the general population. Interestingly, all the values reported in Marsh and Hau (2003) and Seaton et al. (2009) were smaller than those reported in this study for the same country. The magnitude differences might have been results of modelling math self-concept as a latent variable at both levels and of using implicit group mean centring (i.e. reduced measurement and sampling errors). It might also have been a result of using different cross-national databases (i.e. PISA vs. TIMSS) and different populations (15-year-olds vs. eighth graders).

When *perceived* relative standing was added to the within level, the within-level effect of math ability on math self-concept decreased. In fact, in most countries, the effect of perceived relative standing was larger than that of math ability. This is consistent with the prediction from social comparison theory that what matters most is how students perceive, instead of, objective and forced relative standing in class. The largest perceived relative standing effects were observed in the USA (1.27), the Russian Federation (1.27), and Morocco (1.26) and the smallest effects were in Palestine (.35) and Indonesia (.36). However, in contrast to earlier finding that the BFLPE could be eliminated by including perceived relative standing as an additional predictor (Huguet et al., 2009), the BFLPEs found before did not disappear with this additional predictor. In fact, the BFLPE either decreased or maintained at a similar magnitude level.

One limitation of this study is the measure of math self-concept. TIMSS itself was not designed to include a math self-concept measure. Only three items from the student questionnaire seems to capture aspects of this construct. Although math self-concept is used as the construct name, it is not clear whether this construct and this measure are universal across countries. Seemingly similar statements may bear different meanings and interpretations under different cultural value systems. For example, in cultures where being modest is considered a virtue, responses to items such as 'I usually do well in mathematics' likely not only reflect a person's perceptions of his/her math ability but also indicate the person's conformity to the cultural value. A similar finding was regarding the reading self-concept construct in a study where students from Hong Kong and the UK were compared (Wang, Wang, & Osterlind, 2011). In addition, translating the items into different languages during TIMSS 2007 data collection may alter the original meaning carried in the English language. As Murphy and Alexander (2000) have pointed out, most motivation constructs developed represent a Western philosophical orientation. And the majority of cross-cultural research begins with an American instrument and the next step is to translate it into the language where the study is to be conducted (Tirri & Campbell, 2010). Future research is needed to examine representational differences in constructs as well as measurement differences.

It is not clear whether students had a choice to be in a high-achieving or low-achieving class in the 49 countries studied. Based on findings in the present study, being in a high-achieving class in the countries with a significant BFLPE seems detrimental to math self-concept. However, the assimilation effect of being in a high-achieving class may work differently for those who had a choice and those who did not have a choice. Future BFLPE research could study the effect of choice.

Given the fact that math ability/achievement is a cognitive assessment about individuals, it is likely that the BFLPE does not generalise to situations when assessments are for groups. In fact, the finding in earlier studies that BFLPE is specific to academic subjects is likely due to the 'individual' aspect of academic assessment. It will be interesting to examine people's self-perceptions when group assessments are used.

One challenge in large-scale cross-national research is the shear amount of data and number of analyses, especially when results are reported for separate countries. This study is an example of how data and analyses can be handled across different software packages. For example, Mplus is an ideal package to run the models and to output desired statistics. However, it is relatively slow when many parallel runs have to be carried out. The MONTECARLO option in Mplus may be used but later data processing would require a lot of work. On the other hand, more general software such as SAS and R has the advantage of faster, more flexible, and more efficient data-processing capacity. However, since they are not particularly designed for structural equation modelling, many usual statistics are not produced by default. Fortunately, the 'MplusAutomation' package in R developed by Hallquist (2011) can be used to call Mplus so that the advantage of both programmes can be utilised. With some data management skills, data can also be transferred between different platforms.

References

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

Asparouhov, T. (2004). Weighting for unequal probability of selection in multilevel modeling. *Mplus Web Notes: No. 8*. Retrieved from http://www.statmodel.com/download/webnotes/MplusNote81.pdf

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, *35*, 439–460. doi:10.1080/03610920500476598

Asparouhov, T., & Muthen, B. (2006). *Multilevel modeling of complex survey data*. Paper presented at the American Statistical Association, Seattle, WA.

Bong, M., & Skaalvik, E. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*(1), 1–40. doi:10.1023/A:1021302408382

Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation*. Washington, DC: American Psychological Association. doi:10.1037/10197-000

Chiu, M.-S. (2012). The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology*, *104*, 87–107.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

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

Espenshade, T. J., Hale, L. E., & Chung, C. Y. (2005). The frog pond revisited: High school academic context, class rank, and elite college admission. *Sociology of Education*, *78*, 269–293. doi:10.1177/003804070507800401

Festinger, L. (1954). A theory of social comparison processes. *Human relations*, *7*, 117–140. doi:10.1177/001872675400700202

Foy, P., & Olson, J. F. (Eds.) (2009). *TIMSS 2007 user guide for the international database*. Boston, MA: TIMSS & PIRLS International Study Center.

Graham, J., Olchowski, A., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. doi:10.1007/s11121-007-0070-9

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

Hay, I., Ashman, A., & van Kraayenoord, C. E. (1997). Investigating the influence of achievement on self-concept using an intra-class design and a comparison of the PASS and SDQ-1 self-concept tests. *British Journal of Educational Psychology*, *67*, 311–321. doi:10.1111/bjep.1997.67.issue-3

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., … Nezlek, J. (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. doi:10.1037/a0015558

Kaplan, D., & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling*, *6*, 305–321. doi:10.1080/10705519909540138

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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. doi:10.1037/a0012869

Marsh, H. W. (1986). Global self-esteem: Its relation to specific facets of self-concept and their importance. *Journal of Personality and Social Psychology*, *51*, 1224–1236. doi:10.1037/0022-3514.51.6.1224

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

Marsh, H. W. (1987b). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. *Journal of Educational Measurement*, *24*, 17–39. doi:10.1111/jedm.1987.24.issue-1

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

Marsh, H. W. (1990b). *Self-description questinnaire-II*. San Antonio, TX: The Psychological Corporation.
Marsh, H. W. (1990c). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, *82*, 623–636. doi:10.1037/0022-0663.82.4.623
Marsh, H. W. (1992). *Self description questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. A test manual and research monograph*. Macarthur: University of Western Sydney, Faculty of Education.
Marsh, H. W., Chessor, D., Craven, R., & Roche, L. (1995). The effect of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, *32*, 285–319. doi:10.3102/00028312032002285
Marsh, H. W., & Hau, K.-T. (2003). Big-fish–little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*, 364–376. doi:10.1037/0003-066X.58.5.364
Marsh, H. W., 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. doi:10.1037/0022-3514.78.2.337
Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764–802. doi:10.1080/00273170903333665
Marsh, H. W., & O'Neill, R. (1984). Self description questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, *21*, 153–174. doi:10.1111/jedm.1984.21.issue-2
Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*, 319–350. doi:10.1007/s10648-008-9075-6
Marsh, H. W., Trautwein, U., Lüdtke, O., & Koller, O. (2008). 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*, 510–524. doi:10.1037/0022-0663.100.3.510
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, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. doi:10.1007/BF02294457
Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology*, *25*, 3–53. doi:10.1006/ceps.1999.1019
Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American.
Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: The Big-fish-little-pond-effect for PISA 2006. *Educational Psychology*, *31*, 629–656. doi:10.1080/01443410.2011.586416
Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "×" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22*, 1058–1066. doi:10.1177/0956797611415540
Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*, 193–203. doi:10.1037/0022-0663.86.2.193
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rogers, C. M., Smith, M. D., & Coleman, J. M. (1978). Social comparison in the classroom: The relationship between academic achievement and self-concept. *Journal of Educational Psychology*, *70*, 50–57. doi:10.1037/0022-0663.70.1.50
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/SERIES1345
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall. doi:10.1201/CHMONSTAAPP
Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology*, *101*, 403–419. doi:10.1037/a0013838
Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407–441. doi:10.3102/00346543046003407
Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, *9*, 475–502. doi:10.1207/S15328007SEM0904_2
Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling*, *13*, 28–58. doi:10.1207/s15328007sem1301_2
Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science*, *11*, 159–163. doi:10.1111/cdir.2002.11.issue-5
Thijs, J., Verkuyten, M., & Helmond, P. (2010). A further examination of the big-fish–little-pond effect. *Sociology of Education*, *83*, 333–345. doi:10.1177/0038040710383521
Tirri, K., & Campbell, J. R. (2010). Current trends and dilemmas in cross-cultural research. In D. K. Sharpes (Ed.), *Handbook on international studies in education* (pp. 17–29). Charlotte, NC: Information Age.
Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen, & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). Slough: NFER.
von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics*, *32*, 233–251. doi:10.3102/1076998607300422
Wang, Q., Wang, Z., & Osterlind, S. J. (2011). Modeling the effects of home and student factors on text comprehension. *New Waves - Educational Research & Development*, *14*, 22–33.
Wang, Z., Osterlind, S. J., & Bergin, D. A. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *International Journal of Science and Mathematics Education*, *10*, 1215–1242.
Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128. doi:10.1016/j.stueduc.2005.05.005
Zeidner, M., & Schleyer, E. J. (1999). The big-fish–little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology*, *24*, 305–329. doi:10.1006/ceps.1998.0985