Article

Educational and Psychological Measurement
73(4) 601–630

DOI: 10.1177/0013164413481802
epm.sagepub.com
SAGE

Do They Feel the Same Way About Math? Testing Measurement Invariance of the PISA "Students' Approaches to Learning" Instrument Across Immigrant Groups Within Germany

Micha Segeritz[1] and Hans Anand Pant[2]

Abstract

This article summarizes the key finding of a study that (a) tests the measurement invariance (MI) of the popular Students' Approaches to Learning instrument (Programme for International Student Assessment [PISA]) across ethnic/cultural groups within a country and (b) discusses implications for research focusing on the role of affective measures in immigrant and minority education. The Students' Approaches to Learning instrument captures some of the most prominent constructs in educational psychology. Results indicate significant variation in MI across various affective scales and across cultural groups. This study demonstrates that even if MI for specific scales is established across countries, it is still necessary to test MI across cultural groups within a country. We then discuss implications of MI across immigrant groups and highlight the relevance of MI testing for all research studying the affective conditions of educational achievement among immigrant students or the educational motivation of minority students.

[1]New York University, New York, NY, USA
[2]Humboldt Universität zu Berlin, Berlin, Germany

Corresponding Author:
Micha Segeritz, New York University, 285 Mercer Street, 3rd Floor, New York, NY 10003, USA.
Email: michael.segeritz@nyu.edu

Keywords

educational tests and measurements, achievement motivation, psychological tests, immigrant students, measurement invariance

The Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development [OECD], 2001)[1] is the largest educational study worldwide, and its publicly available data sets provide unique opportunities for immigration, social, and behavioral researchers. The inclusion of a large battery of psychological constructs allows researchers to connect the achievements of immigrant groups to affective measures such as instrumental motivation or domain-specific self-concept and investigate popular theoretical concepts such as ''immigrant optimism'' or ''acculturative stress'' (e.g., Kao & Tienda, 1995; Roche & Kuperminc, 2012). In order to yield valid interpretations, the instruments have to be understood equivalently and answered with the same reference frame across groups. However, the notion that affective measures can easily be compared across cultural groups is highly contested within cross-cultural psychology (Berry, Poortinga, Segall, & Dasen, 2002). A large body of research suggests that cultural aspects have strong impacts on motivation, emotion as well as perception, and threaten potentially the generalizability and universality of psychological constructs across groups (e.g., Berry et al., 2002; Markus & Kitayama, 2010). Therefore, the psychometric literature broadly agrees that researchers should only compare scale scores across cultural groups after establishing measurement invariance (e.g., Byrne & Van de Vijver, 2010; Van de Vijver & Leung, 1997; Van de Vijver & Poortinga, 2002). The definition of measurement invariance (MI) states that the scale score only depends on the underlying latent construct and is independent of group membership (Meredith, 1993). A wide range of research investigates the cultural influence on affective measures between cultural groups residing in different countries; however, less attention is directed toward establishing MI for (cultural) groups within one country (Avery, Tonidandel, Thomas, Johnson, & Mack, 2007; Marcoulides, Emrich, & Marcoulides, 2008; Shavelson & Byrne, 1987).

The purpose of this article is to investigate the important, but often overlooked, question of establishing MI for within-country immigrant groups using data from the German PISA-2003 database. Even though several publications analyzed the psychometric quality of affective and motivational PISA scales (e.g., OECD, 2005), to our knowledge, there is no study testing MI for PISA's ''Students' Approaches to Learning (SAL)'' instrument. The SAL instrument is an item battery (nine scales, 45 items) that measures some, of the most widely used constructs in educational psychology. The oversampling of immigrant students in the German PISA-2003 sample allows to test MI for students with a German, Turkish, and former USSR origin.

Students' Approaches to Learning

Self-regulated learning theories describe complex and multidimensional learning processes (e.g., Marsh, Hau, Artelt, Baumert, & Peschar, 2006; Sitzmann & Ely, 2011). They integrate various dimensions such as goal setting, motivation, and learning strategies and self-related cognition to understand how and why individuals engage in learning. Self-regulated learning models provide overarching heuristics that include established constructs and offer a perspective to understand learning processes both in institutional and autonomous situations (Sitzmann & Ely, 2011). Guided by Boekaert's (1997) model of self-regulated learning, the OECD developed the multifactor SAL instrument that aims at measuring prerequisites for self-regulated learning processes and motivation related to education (OECD, 2005).

The PISA-2003 SAL instrument attempts to measure three dimensions of self-regulated learning: (a) motivational preferences, (b) self-related cognitions, and (c) learning strategies (Marsh, Hau, et al., 2006). Additionally, the SAL instrument is complemented by two scales measuring learning preferences of students (d). These scales do not directly capture aspects of self-regulated learning, but rather preferences of students for specific settings in which learning occurs (Marsh, Craven, Hinkley, & Debus, 2003). Because mathematics was of key interest in the PISA-2003 study, the SAL instrument has a domain-specific focus on mathematics (for an overview, see Table A1 in the appendix). Methodological reviews have established the psychometric properties of the individual scales (OECD, 2005) and Marsh, Hau, et al. (2006) demonstrated that the factor structure of an earlier version of the SAL instrument was well defined. In the remainder of this section, we describe the different dimensions and factors of the SAL instrument.

Motivational preferences related to mathematics (a) were captured by the scales ''instrumental motivation in mathematics'' and ''interest in and enjoyment of mathematics.'' ''Instrumental motivation in mathematics'' refers to extrinsic motivation that has proven to be an important predictor of course selection and performance (Marsh, Hau, et al., 2006). Interest in mathematics, on the other hand, corresponds with intrinsic motivation in R. M. Ryan and Deci's (2000) learning theory. Previous research showed that interest in a subject correlates positively with the intensity and persistence of learning efforts as well as with choosing more elaborate learning strategies (Silvia, 2006).

The two factors capturing *self-related cognitions regarding mathematics (b)* are ''self-concept in mathematics'' and ''self-efficacy in mathematics.'' Self-concept refers to broader domain-specific cognitive abilities that individuals attribute to themselves (Marsh & O'Mara, 2008) whereas ''self-efficacy'' refers to the level of confidence to accomplish and solve specific tasks and problems in certain areas (Pajares & Miller, 1994). Previous research has demonstrated that positive self-related cognition is connected to ambitious goal selection, high effort in pursuing these goals, and positive domain-specific outcomes such as achievement and well-being (Caprara, Veccione, Alessandri, Gerbino, & Barbaranelli, 2011; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006). The positive outcomes in turn may

increase domain specific motivation and suggests reciprocity between self-related cognition and motivation.

A central aspect of self-regulated learning are *learning strategies (c)* that help individuals plan learning processes and accomplish the set goals in an effective and efficient manner (e.g., Pintrich, 2000; Zimmerman, 2000). The PISA-2003 SAL instrument incorporates ''memorization strategies,'' ''elaboration strategies,'' and ''control strategies.'' The first two constructs are primarily cognitive strategies including specific techniques that help individuals consolidate their acquired knowledge and gain a deeper understanding of study material. Control strategies are rather metacognitive and monitor the learning process and make adjustments as necessary (Zimmerman, 2000).

Although not part of the self-regulated learning framework, *students' preferences regarding learning situations (d)* is a useful extension to better understand self-regulated learning (Marsh et al., 2003; Marsh, Hau, et al., 2006). These two factors, ''preference for co-operative learning'' and ''preference for competitive learning'' may provide relevant insight into learning motives and preferred learning situations. Both, collaborating in groups and learning independently (i.e. competitive learning) are important competencies that are not necessarily mutually exclusive.

Overall, a large body of research indicates that motivation to learn, self-related cognition, and the appropriate use of learning strategies may be crucial for successful learning processes (Marsh, Hau, et al., 2006; Sitzmann & Ely, 2011; Wigfield, Eccles, & Rodriguez, 1998).

Cultural Differences in Affective Measures and Measurement Invariance

The different components of self-regulated learning may help better understand differences in educational processes and trajectories between native and immigrant students (e.g., Christensen & Segeritz, 2008; Stanat & Christensen, 2006; Stanat, Segeritz, & Christensen, 2010). However, theoretical work and empirical analysis suggest that psychological instruments are not necessarily measurement invariant across cultural and other social groups. Resulting scores may therefore not yield valid comparisons across groups (e.g., Avery et al., 2007; Byrne & Van de Vijver, 2010; Mylonas, 2009; Van de Vijver & Leung, 2011).

The notion of generalizability of psychological constructs and scores across cultural groups such as immigrant groups is controversially discussed within cross-cultural psychology and related domains (e.g., Fontaine, 2011; Van de Vijver & Leung, 2011). It touches fundamental questions of comparability of meaning and understanding of human behavior across cultural groups (Berry et al., 2002). Markus and Kitayama (1991), for example, describe the different framing of ''self'' between certain more collectivistic Asian cultures and more individualistic Western cultures and conclude that these differences affect motivation, cognition, and emotion. Self-concept questionnaires developed for Western individuals therefore may not capture

the wider meanings of ''self'' in collectivistic cultures and self-concept may not be construct equivalent across certain groups.

Moreover, when comparing scale scores across groups, subjects do not only have to apprehend the underlying construct and each item in the same way, but they also have to apply the same rating reference frame across groups (e.g., Van de Vijver & Leung, 2011). In other words, comparable ''true scores'' on underlying constructs have to translate to the same scale scores across groups and a one unit increase in the underlying latent factor has to correspond with equivalent changes on related items. Previous studies that question these underlying assumptions have reported response styles that differ by culture and methodological artifacts related to interactions between group membership and measurement scales (Johnson, Kulesa, Llc, Cho, & Shavitt, 2005; Twenge & Crocker, 2002).

Thus, sociocultural group membership may affect the conceptualization and meaning of psychological constructs, the units of scales, as well as the score of scales (Byrne & Van de Vijver, 2010). Confirmatory factor analysis (CFA) provides a framework to investigate these threats to valid group comparisons and to test them in a series of sequentially more restrictive models to establish MI (e.g., Byrne, 2008; Vandenberg & Lance, 2000).

Testing Psychological Constructs for Measurement Invariance

When social scientists try to capture psychological constructs with survey instruments, they assume that each subject has an unobserved true latent variable score η, representing the respective latent construct (e.g., T. D. Little, 1997; Wu, Li, & Zumbo, 2007). The magnitude of η predicts the outcomes of the associated observable questionnaire items. In a CFA, for each item j an equation is specified that predicts the item score y as a function of the unobserved latent variable η. In multigroup CFAs (MG-CFAs), the models simultaneously estimate the mean and covariance structure of subjects with different group memberships (T. D. Little, 1997). The mean and covariance structures analysis (MACS) extends the often used covariance analysis for multigroup comparisons by incorporating the means of the items. MACS models can be expressed by a series of equations with the following form:

$$y_{ijg} = \tau_{jg} + \lambda_{jg}\eta_{ig} + r_{ijg} \quad (1)$$

Equation (1) estimates the observed item score y of item j for person i in group g based on an underlying factor (i.e., latent variable/construct) η_{ig}.[2] The equation states that the item score y_{ijg} is a linear combination of an item-specific intercept τ_{jg}, the respective factor score η_{ig} times the factor loading λ_{jg}, and a random error component r_{ijg}. The item intercept τ_{jg} indicates the respective item score when the factor score of the latent variable is 0. Since τ_{jg} is considered constant for all subjects in the same group, it does not have a subscript i. This is also true for λ_{jg}, the factor loading. λ_{jg} indicates the change in the item score y_{ijg} when the underlying factor score η_{ig} of a person i in group g changes by one unit. The subscript g in this formula indicates

that the parameter estimates τ and λ are allowed to vary across groups and, therefore, depend on group membership. However, if an underlying construct is comparable across groups and if it can be measured comparably across groups, then Equation (1) should be independent of group membership (e.g., Wicherts & Dolan, 2010). This means, in each group, a specific latent factor score is associated with the same item scores, independent of the group membership. This requirement is captured in the statistical definition of MI by Meredith (1993) which states that each item score is only a function of the underlying factor η but not of group membership.

The statistical approach most commonly used for establishing MI tests a set of necessary conditions in a hierarchical stepwise procedure (e.g., Byrne, 2008; Vandenberg & Lance, 2000). In each step, a set of parameters is fixed to be equal across groups and the fit of the resulting model is compared with the previous model's fit. Full MI can be assumed when the completely restricted model fits the data similarly well as the intiial model in which all parameters were estimated freely. The three necessary steps for testing MI are (a) configural invariance, (b) metric invariance (weak invariance), and (c) scalar invariance (strong invariance; e.g., Byrne & Van de Vijver, 2010).

The first step in testing MI is to establish *configural invariance* (Horn & McArdle, 1992). Configural invariance states that the factor structure and loading pattern are the same across groups. To test configural invariance, the hypothesized factor structure (i.e., Table A1 in the appendix) is specified for each group in the same way but without constraining the estimation of the specified parameters.[3] Thus, the factor loadings, λ, error variances, r, and item intercepts, τ, are simultaneously and independently estimated for each group. A good model fit (see discussion of fit indices below) indicates that members of each group apprehend the constructs conceptually in a similar way, and therefore the instrument(s) are configurally invariant.

Metric invariance postulates that a one-unit change in the latent factor score has the same meaning across groups (Vandenberg & Lance, 2000). According to Formula (1), a one-unit change in the factor score, η, results in the same average change across groups for each item measuring aspects of the underlying construct. This would mean that λ_{jg} is the same for all groups and does not depend on group membership (i.e., $\lambda_{j,1} = \lambda_{j,2}$, where subscripts 1 and 2 indicate members of Group [g] 1 and Group 2, respectively). In order to test metric invariance, we estimate an MG-CFA that constrains all factors loadings of like items to be equal across groups (Byrne, 2008). If this model fits the data equally well as the previous configural invariance model, we can conclude that the units are equivalent across groups and can proceed to test scalar invariance.

Scalar invariance states that a given latent variable score results in each group in the same item and scale scores (Van de Vijver & Leung, 1997). For example, if students from different groups are equally motivated and thus have the same latent variable scores, they show on average equal scores on the same items and also equal scale scores—independent of group membership. Thus, to test scalar MI, we test the equality of the mean structure in addition to the equality of the factor loadings in the

previous step. More precisely, the factor scores, η, are arbitrarily fixed to a specific value (often fixing all factor scores to 0), and the intercepts, τ, are constrained to be equal across groups. If the factor scores are fixed to the same value across group and the construct(s) are scalar invariant, then each item is expected to have the same intercept across groups (e.g., $\tau_{j,1} = \tau_{j,2}$, where the subscripts 1 and 2, indicate members of Group 1 and Group 2, respectively). As described above, if this more restrictive model fits the data as well as the previous model (i.e., no substantial decrement in the model fit) where only the factor loadings were constrained to be equal across groups, scalar invariance is established (Beckstead, Yang, & Lengacher, 2008). Consequently, given that the error variances are conditionally independent (see, discussion of strict invariance below), the factor means can be justifiably compared across groups (Byrne & Van de Vijver, 2010).

Data and Method

PISA Data Set

The German PISA-2003 sample is particularly appropriate to test MI of the SAL instrument across immigrant groups. Germany has a sizable immigrant population, and policy makers opted for an optional national extension that oversampled immigrant students and therefore enables comparisons across immigrant groups. In our sample, we included students who themselves or their parents were born in either the former USSR or Turkey. Additionally, we included students without an immigrant background. Thus, the sample includes 2,428 students with a Turkish background, 2,948 students with a former USSR background and 27,670 students with no immigrant background (see Table 1). Students with a former USSR or a Turkish background are the two largest immigrant groups in Germany with about 4.8% and 4.1% of the total student population, respectively (Segeritz, Walter, & Stanat, 2010). Since the vast majority of immigrants from the former USSR are of German heritage, it seems appropriate to cluster them into one group (Takle, 2011).

Students' Approaches to Learning Instrument

The multifaceted SAL instrument with 9 factors and 45 items was compiled and developed by an OECD-led expert group (for an extensive documentation of the development process of the SAL instrument and the individual scales, see Marsh, Hau, et al., 2006; Peschar et al., 1999a, 1999b). Each scale is measured with four to eight items and the internal consistency seems acceptable to good for the majority of scales (see Table 2). The inclusion of ''memorization strategies'' seems justifiable since acceptable internal consistency was just marginally missed and the low Cronbach's alpha was likely at least partially caused by the low number of items in the scale.

Table 1. Sample Sizes for Complete Sample and Subsamples.

		N	Percentage
Complete sample	Native students	27,670	83.7
	USSR background	2,948	8.9
	Turkish background	2,428	7.3
	Total	33,046	100.0
Subsample 1	Native students	13,915	83.7
	USSR background	1,491	9.0
	Turkish background	1,219	7.3
	Total	16,625	100.0
Subsample 2	Native students	13,755	83.8
	USSR background	1,457	8.9
	Turkish background	1,209	7.4
	Total	16,421	100.0

Table 2. Descriptive Statistics and Cronbach's Alpha.

Variable	No. of items	Cronbach's α	*N*	Mean	*SD*
INSTMOT—Instrumental motivation to learn mathematics	4	.82	32,901	2.89	0.71
INTMAT—Interest in and enjoyment of mathematics	4	.90	32,909	2.32	0.84
MATHEFF—Mathematics self-efficacy	8	.81	32,844	3.07	0.56
SCMAT—Mathematics self-concept	5	.91	32,805	2.55	0.82
MEMOR—Memorization/ rehearsal strategies	4	.68	32,709	2.55	0.64
ELAB—Elaboration strategies	5	.74	32,744	2.36	0.60
CSTRAT—Control strategies	5	.72	32,813	3.10	0.54
COOPLRN—Preference for cooperative learning situations	5	.76	32,734	2.71	0.62
COMPLRN—Preference for competitive learning situations	5	.84	32,793	2.56	0.69

Method

Before testing the three steps of MI, we first evaluate in a preliminary step whether the proposed factor structure fits the data of each student group (Step 1; Byrne, 2008). Then, an MG-CFA model estimates simultaneously, but independently, the proposed factor structure across the three groups (Step 2). An appropriate baseline model fit indicates configural invariance and is a prerequisite for continuing the

invariance testing procedure. A well-fitting baseline model will be used to evaluate the model fit of increasingly more restrictive models in the following steps. That is, in subsequent nested models, the factor loadings (metric invariance) and item intercepts (scalar invariance) are constrained to be equal across groups (Steps 3 and 4). If the model fit does not decrease substantially compared with a less restrictive model, it is assumed that the more restrictive model fits the data equally well and that the respective level of MI is achieved (Byrne, 2008; Van de Vijver & Leung, 1997).

All model parameters are estimated using a maximum likelihood estimator. A sandwich-type covariance matrix is used to calculate standard errors and chi-square statistics that are robust to nonnormality of the data (maximum likelihood robust [MLR]; Yuan & Bentler, 2000). All available information of observations with partially missing data are included and analyzed under the ''missing at random'' assumption (R. J. A. Little & Rubin, 2002). Under this assumption, missing values are allowed to depend on observed but not on unobserved variables (Graham, 2009; Yuan & Bentler, 2000). The overall absolute model fit of the different models is evaluated via the incremental comparative fit index (CFI) and the root mean square error of approximation (RMSEA). As is common practice, a CFI value greater than .90 and RMSEA values less than .08 indicate an acceptable model fit, and a CFI value greater than .95 and an RMSEA value less than .05 indicate a good model fit (Browne & Cudeck, 2003).

Traditionally, likelihood ratio tests based on the chi-square distribution are used to evaluate if a more complex model, such as a model with more equality constraints, fits the data as well as the model in which it is nested (Vandenberg & Lance, 2000). The difference in the chi-square statistics, $\Delta\chi^2$, between the two models is evaluated for statistical significance. However, various studies have shown that likelihood ratio tests are overly sensitive to increasing sample sizes and pick up even trivial differences between groups (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Thus, because of our large sample sizes (roughly between 1,200 and 14,000 students per group), we are mainly using alternative fit indices to test MI. Even though this approach is lacking a clear criterion such as statistical significance, it has been proven to be superior with large samples compared with conventional likelihood-ratio test procedures (Meade et al., 2008; Vandenberg & Lance, 2000). Meade et al. (2008) suggest that a change in the CFI (ΔCFI) of up to .002 is acceptable and indicates that the decrease in model fit can be considered trivial. Although the CFI is less sensitive to sample size, Monte Carlo studies have shown that with increasing sample sizes random nonsubstantial differences may be picked up as well (Meade et al., 2008). Therefore, because of the large sample size of the study, comparisons of the the group-specific parameter estimates are incorporated in the decision process (A. M. Ryan, West, & Carr, 2003).

During the process of finding a baseline model and conducting the sequential MI testing procedures, occasionally post hoc model specifications, suggested by modification indices, are made if they seemed theoretically justified. To confirm that these partly data-driven adjustments were not solely because of idiosyncrasies in the data

set, this study used a split-half cross-validation approach (Byrne et al., 1989). A randomly selected subsample (No. 1) underwent the MI testing procedures described above, including the post hoc model respecification. Afterward, Subsample 2 was used to evaluate whether the respecifications were warranted and could be replicated. The two subsamples were partitioned based on a random number algorithm and contain roughly 16,500 students, respectively (see Table 1).

Results

Step 1: The Proposed Model Was Fitted on Each Group Independently to Ensure Good Model Fit

Following Byrne's (2008) suggestion, we initially tested the hypothesized nine-factor structure separately for German students and students with a Turkish or former USSR background. Based on the chi-square statistic and the alternative fit index CFI, the initial model did not fit the data of the randomly selected subsample well for any of the three groups (Table 3: Model 1a [German/nonimmigrant background], Model 1c [former USSR background], and Model 1e [Turkish background]). Exploration of the modification indices connected the low model fit to three variables that did not conform to the imposed factor structure. We excluded these variables from the following models (see footnote to Table A1 in the appendix). Additionally, modification indices indicated that the model fit would improve if a set of residual covariances were estimated freely. Closer inspection revealed that five of the suggested item pairs were very similar to each other and, therefore, provided a theoretical justification for freely estimating their residual covariances (Avery et al., 2007). For example, unlike the other items measuring self-efficacy in mathematics, the Items ST31Q05 and ST31Q07 explicitly included mathematical equations, which rendered them more similar to each other than to other related items. The chi-square statistic improved strongly, but was still statistically significant while both alternative fit indices RMSEA and CFI indicate that the data fit the model reasonably well (Model 1b [German], Model 1d [former USSR], and Model 1f [Turkish]). Given the very large sample size and the well-known sensitivity of the chi-square statistic to sample size, detecting statistically significant model misfit was not surprising (Hu & Bentler, 1999; Marsh, Hau, et al., 2006). Therefore, similar to other studies with large sample sizes (e.g., Avery et al., 2007), we primarily focused on the alternative fit indices RMSEA and CFI that indicated acceptable fit and have proven to perform well with large sample sizes.

Step 2: Estimating the Proposed Baseline Model Simultaneously on All Groups

In this step, a baseline model was estimated that combined the data of all three groups and estimated the model parameters independently for each group (Model 2).

Table 3. Multigroup Confirmatory Factor Analysis for Subsample 1.

		Model	MLR-χ	*df*	CFI	TLI	RMSEA	SRMS	*N*	Items	Factors
Subsample 1											
Group-specific baseline models	Students with German background (G)	1a	13058.28	909	.887	.877	.031	.064	13,909	45	9
		1b	7,921	778	.929	.921	.026	.05	13,909	42	9
	Students with former-USSR background (f-U)	1c	2156.66	909	.881	.87	.03	.073	1,487	45	9
		1d	1472.116	778	.93	.922	.024	.055	1,487	42	9
	Students with Turkish background (T)	1e	1929.245	909	.876	.865	.03	.066	1,218	45	9
		1f	1481.576	778	.907	.897	.027	.059	1,218	42	9
Configural invariance	Parameters freely estimated across groups	2	10201.41	2,334	.926	.918	.025	.051	16,614	42	9
Metric invariance	Factor loadings constrained across groups	3	10348.85	2,400	.925	.92	.024	.052	16,614	42	9
Scalar invariance	Intercepts constrained across all groups	4a	11280.04	2,484	.917	.914	.025	.058	16,614	42	9
	Intercepts constrained across G and f-U students	4b	10751.93	2,442	.922	.918	.025	.054	16,614	42	9

(continued)

Table 3. (continued)

	Model	MLR-χ	*df*	CFI	TLI	RMSEA	SRMS	*N*	Items	Factors
Intercepts constrained across G and f-U students	4c	10588.95	2,429	.923	.919	.025	.053	16,614	42	9
Intercepts constrained across G and T students	4d	10928.55	2,442	.92	.916	.025	.057	16,614	42	9
Intercepts constrained across G and T students	4e	10769.09	2,429	.922	.917	.025	.056	16,614	42	9
Intercepts constrained across G and T students	4f	10454.53	2,412	.925	.919	.025	.054	16,614	42	9
Intercepts constrained across all groups	4g	10689.96	2,441	.923	.918	.025	.054	16,614	42	9

Note. G = German; f-U = former USSR; T = Turkish; MLR = maximum likelihood robust; CFI = comparative fit index; RMSEA = root mean square error of approximation; TLI = Tucker–Lewis index; SRMS = standardized root mean square residual. In Models 4c and 4e, only intercepts of MATHEFF, MEMOR, ELAB, CSTRAT, COOPLRN, and COMPLRN constrained across groups. In Model 4f, only intercepts of MEMOR, CSTRAT, and COOPLRN constrained across groups. Model 4g applies the constraints of Model 4c for students with a former USSR background and of Model 4f for students with a Turkish background.

The acceptable model fit (CFI = .926, RMSEA = .025) indicated that the factor structure and loading patterns were similar across groups. Thus, configural invariance, the prerequisite for more stringent tests of MI, can be assumed and the configural invariance model can be used as the baseline model for the more restrictive model below (Byrne, 2008).

Step 3: Testing the Factor Loading Equality

Factor loading equality or metric invariance was tested by constraining the factor loadings to be equal across groups (Model 3: CFI = .925, RMSEA = .024). The ΔCFI of .001 between Model 2 and Model 3 indicated that the factor loadings were not substantially different across groups. Therefore, one can assume that the measurement units are the same across groups. A one-unit change in the underlying latent variable resulted in equal changes in like items for students with a Turkish, former USSR, or German background.

Step 4: Testing Equivalence of Intercepts Across Groups

Subsequently, scalar MI was tested. In this model, for identification reasons, the factor scores were fixed to 0 for all factors across all groups. Thus, the estimated item intercepts indicated the respective item score when the related latent variable score was 0. To estimate scalar MI, the item intercepts of like items must be the same across groups and were therefore constrained to be equal across groups. The ΔCFI of .008 between Model 3 and Model 4a indicated that at least for some factors, the same latent variable score did not result in equivalent item score across student groups. To investigate the lack of scalar MI further, MI was tested separately between German students and students with a former USSR background (Models 4b and 4c) as well as German students and students with a Turkish background (Models 4d, 4e, and 4f).

Model 4b tested MI between students with a former USSR and German background while freely estimating the intercepts for students with a Turkish background. The ΔCFI of .003 compared with Model 3 suggested lack of scalar MI between students with a former USSR and German background for at least some factors. After visually evaluating the freely estimated intercepts in Model 3, it appeared that students with a background from the former USSR tended to rate items related to ''instrumental motivation in mathematics,'' ''interest and enjoyment in mathematics,'' and ''self-concept in mathematics'' higher even if there was no difference in the latent true factor score. This means that students from the former USSR would rate items higher even if they have comparable attitudes on the latent constructs. Thus, the scaled scores between students with a former USSR background and German background would not be comparable. Therefore, in Model 4c, the item intercepts related to these three factors were freely estimated in both groups. Given the ΔCFI of .002 in comparison with Model 3, we concluded scalar MI for all scales except ''instrumental motivation in mathematics,'' ''interest and enjoyment in

mathematics,'' and ''self-concept in mathematics'' for students with a German and former USSR background.

Similarly, we tested invariance of item intercepts between German students and students with a Turkish background by fixing the latent underlying factor to 0 in both groups. Model 4d indicated lack of MI between the two groups (ΔCFI = .005). Visual inspection of the freely estimated item intercepts in Model 3 showed that students with a Turkish background seemed to choose higher scores for items related to ''instrumental motivation in mathematics,'' ''interest and enjoyment in mathematics,'' ''self-concept in mathematics,'' ''elaboration strategies,'' and ''competitive learning'' even when the underlying factor score was equivalent to the factor score of German students. Moreover, the item intercepts of ''mathematics self-efficacy'' indicated lower rating tendencies of students with a Turkish background. Model 4e (ΔCFI = .003) and Model 4f (ΔCFI = .000) showed that only after freeing the estimation procedure for intercepts related to ''elaboration strategies,'' ''competitive learning,'' and ''mathematics self-efficacy'' in addition to the three factors ''instrumental motivation in mathematics,'' ''interest and enjoyment in mathematics,'' and ''self-concept in mathematics'' was scalar MI established.

Next, in Model 4g, we simultaneously estimated the respective group-specific model specifications found in Models 4c and 4f. More precisely, we freely estimated intercepts related to three factors for students with a former USSR background and related to six factors for students with a Turkish background. The ΔCFI of .002 compared with Model 3 indicated scalar MI for the factors whose intercepts were constrained to be equal across the three groups. Thus, only the scale scores of ''memorization/rehearsal strategies,'' ''control strategies,'' and ''preference for cooperative learning situations'' can be compared across all three groups. However, for analysis including only immigrant students from the former USSR, the scale scores of ''mathematics self-efficacy,'' ''elaboration strategies,'' and ''competitive learning'' can be compared validly, too.

Finally, replicating the analysis with Subsample 2, the analyses confirmed all the previously detected structures of MI and lack of MI (see Table A3 in the appendix). This suggested that the results described above were not driven by idiosyncratic patterns in the data.

Discussion and Conclusion

Our analysis showed that the factorial structure of the SAL instrument is comparable across the German and the two immigrant student groups. However, three items seemed not to work well in any of the groups and therefore should be excluded from calculating factor scores. After excluding the three ill-fitting items, configural invariance and metric MI could be established across three groups. This means that a particular change in the underlying construct results in comparable changes in the item scores across all groups. Therefore, a one-unit scale score change has the same

meaning across all groups (unit equivalence). However, metric invariance does not imply that the actual value of a scale score is comparable across groups as well.

When initially testing scalar MI, the results indicated that scale scores cannot readily be compared across all groups. Further analysis indicated that immigrant students tend to answer particular items more positively, even when the underlying factor score is the same. The lack of scalar MI was connected to the scales ''instrumental motivation,'' ''interest in mathematics,'' and ''self concept in mathematics'' for all groups and to ''elaboration strategies,'' ''competitive learning,'' and ''mathematics self-efficacy'' for the comparison between nonimmigrant students and students with a Turkish background. Thus, only scale scores of ''memorization/rehearsal strategies,'' ''control strategies,'' and ''preference for co-operative learning situations'' can be compared validly across all three groups.

On the other hand, the establishment of metric MI for all scales indicated that the units are comparable across groups. Therefore, depending on the research questions, researchers can still use group-centered scales and conduct analysis that are only concerned with relative differences to the group-specific means (e.g., within-group comparisons of students with high and low scores in instrumental motivation).

Moreover, the absence of scalar MI, and more precisely, the fact that given comparable underlying construct characteristics, students from the former USSR and in particular students with a Turkish background tend to answer questionnaire items more optimistically is a very interesting finding for immigration researchers. This result may shed further light on the often-reported results that immigrant students are reported to have a very positive view toward schooling and their future (immigrant optimism), but at the same time their actual achievement falls short compared with native students (Kao & Tienda, 1995). Potentially the occasionally reported relatively high values on scales such as ''instrumental motivation,'' ''interest in mathematics,'' and ''self-concept in mathematics'' could be a function of stronger positive answer tendencies rather than actual higher motivation or ''immigrant optimism.''

This article demonstrates that it is important to not only establish MI between groups residing in different countries, which currently is the common practice in the measurement literature but also between cultural, racial, and other groups residing in the same country (see also, Avery et al., 2007; Byrne & Watkins, 2003; Mylonas, 2009). The article also highlights two important methodological aspects that are often overlooked in applications of MG-CFAs, specifically the incorporation of the MACS modeling approach (T. D. Little, 1997) and the split-half cross-validation approach (Byrne et al., 1989). Both these methodological approaches significantly strengthen the validity of the results. First, by explicitly incorporating the covariances as well as the means in the testing procedure, our MACS analysis detected variations in answer pattern of immigrant groups that were not connected to meaningful differences in the underlying construct. These noninvariant patterns remain undetected when a testing procedure only focuses on the covariance structures of the measures. Second, similar to many other CFA modeling approaches, this article had to make some post hoc adjustments to the hypothesized underlying measurement

model based on theoretically justifiable, but empirically driven results. Due to the size of the PISA data set, these model specification indices could be derived from a randomly selected split-half subsample. Therefore, the modifications could be tested and validated with the independent second half of the data set and thereby show that the alterations were not driven by idiosyncratic relationship in the data, but by meaningful structures that independently existed in the second half of the data set.

The results of this study also highlight the need for further research in the field of immigrant education and cross-cultural psychology. Particular additional analysis regarding the different scaling of some of the motivational scales and its relation to ''immigrant optimism'' could shed light on the paradox of the regularly reported high motivation and occasionally low performance of immigrant students. Potentially, qualitative approaches, such as the ''think-aloud'' methodology could be particularly helpful (Simon & Ericsson, 1998).

Future research should also focus on generational differences of immigrant groups. According to assumptions of the straight-line assimilation theory, cultural differences should decrease with longer residency and eventually disappear (Alba, 2008). Therefore, one would assume that measurement noninvariance should decrease across generations and potentially disappear for the second or third immigrant generation. However, proponents of the segmented assimilation theory hypothesize that not all immigrant groups will necessarily assimilate to a homogeneous core of the host country (Haller, Portes, & Lynch, 2011). Rather, they may assimilate to a socioeconomically disadvantaged group already residing in the host country for a longer time. Following this path of argumentation that proposes the creation of a culturally distinct social group within a society, MI may not be achieved for certain members of socially disadvantaged immigrant groups—even if they reside in a host country for several generations.

Table A1. Students' Approaches to Learning (SAL) Instrument.

Dimension	Scale/construct		Items	
Motivational preferences	INSTMOT—Instrumental motivation to learn mathematics	Thinking about your views on mathematics: To what extent do you agree with following statements?	ST30q02	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on.
			ST30q05	Learning mathematics is worthwhile for me because it will improve my career (prospects, chances).
			ST30q07	Mathematics is an important subject for me because I need it for what I want to study later on.
			ST30q08	I will learn many things in mathematics that will help me get a job.
	INTMAT—Interest in and enjoyment of mathematics	Thinking about your views on mathematics: To what extent do you agree with following statements?	ST30q01	I enjoy reading about mathematics
			ST30q03	I look forward to my mathematics lessons
			ST30q04	I do mathematics because I enjoy it
			ST30q06	I am interested in the things I learn in mathematics

(continued)

Table A1. (continued)

Dimension	Scale/construct		Items	
Self-related cognitions and beliefs	MATHEFF—Mathematics self-efficacy	How confident do you feel about having to do the following calculations?	ST31q01	Using a (train timetable), how long would it take to get from Zedville to Zedtown?
			ST31q02	Calculating how much cheaper a TV would be after a 30% discount
			ST31q03	Calculating how many square meters of tiles you need to cover a floor
			ST31q04	Understanding graphs presented in newspapers
			ST31q05	Solving an equation like $3x + 5 = 17$
			ST31q06	Finding the actual distance between two places on a map with a 1:10,000 scale
			ST31q07	Solving an equation like $2(x + 3) = (x + 3)(x - 3)$
			ST31q08	Calculating the petrol consumption rate of a car
	SCMAT—Mathematics self-concept	How much do you disagree or agree with the following statements about how you feel when studying mathematics?	ST32q02	I am just not good at mathematics
			ST32q04	I get good (marks) in mathematics

(continued)

Table A1. (continued)

Dimension	Scale/construct		Items	
			ST32q06	I learn mathematics quickly
			ST32q07	I have always believed that mathematics is one of my best subjects
			ST32q09	In my mathematics class, I understand even the most difficult work
Learning strategies	MEMOR—Memorization/ rehearsal strategies	There are different ways of studying mathematics: To what extent do you agree with the following statements?	ST34q06	I go over some problems in mathematics so often that I feel as if I could solve them in my sleep
			ST34q07	When I study for mathematics, I try to learn the answers to problems off by heart
			ST34q09	In order to remember the method for solving a mathematics problem, I go through examples again and again
			ST34q13	To learn mathematics, I try to remember every step in a procedure

(continued)

Table A1. (continued)

Dimension	Scale/construct		Items	
	ELAB—Elaboration strategies	There are different ways of studying mathematics: To what extent do you agree with the following statements?	ST34q02	When I am solving mathematics problems, I often think of new ways to get the answer
			ST34q05	I think how the mathematics problems I have learnt can be used in everyday life
			ST34q08	I try to understand new concepts in mathematics by relating them to things I already know
			ST34q11	When I am solving a mathematics problem, I often think about how the solution might be applied to other interesting questions
			ST34q14	When learning mathematics, I try to relate the work to things I have learnt in other subjects
	CSTRAT—Control strategies	There are different ways of studying mathematics: To what extent do you agree with the following statements?	ST34q01	When I study for a mathematics test, I try to work out what are the most important parts to learn

(continued)

Table A1. (continued)

Dimension	Scale/construct		Items	
			ST34q03	When I study mathematics, I make myself check to see if I remember the work I have already done
			ST34q04	When I study mathematics, I try to figure out which concepts I still have not understood properly
			ST34q10	When I cannot understand something in mathematics, I always search for more information to clarify the problem
			ST34q12	When I study mathematics, I start by working out exactly what I need to learn
Learning preferences in mathematics	COOPLRN—Preference for cooperative learning situations	Thinking about your <mathematics> classes: To what extent do you agree with the following statements?	ST37q02	In mathematics I enjoy working with other students in groups
			ST37q04	When we work on a project in mathematics, I think that it is a good idea to combine the ideas of all the students in a group

(continued)

Table A1. (continued)

Dimension	Scale/construct		Items	
			ST37q06	I do my best work in mathematics when I work with other students
			ST37q08	In mathematics, I enjoy helping others to work well in a group
			ST37q09	In mathematics I learn most when I work with other students in my class
	COMPLRN—Preference for competitive learning situations	Thinking about your <mathematics> classes: To what extent do you agree with the following statements?	ST37q01	I would like to be the best in my class in mathematics
			ST37q03	I try very hard in mathematics because I want to do better in the exams than the others
			ST37q05	I make a real effort in mathematics because I want to be one of the best
			ST37q07	In mathematics I always try to do better than the other students in my class
			ST37q10	I do my best work in mathematics when I try to do better than others

Note. Items ST34q06, ST37q08, and ST37q03 were excluded because of problematic loading patterns. Correlated residuals were allowed between the following item pairs: ST31Q07 and ST31Q05, ST30Q04 and ST30Q03, ST32Q04 and ST32Q02, ST37Q09 and ST37Q02, and ST37Q10 and ST37Q05.

Table A2. Descriptive Statistics and Cronbach's Alpha (Subsamples).

	Variable	Cronbach's α	*N*	Mean	*SD*
Subsample 1	INSTMOT—Instrumental motivation to learn mathematics	.82	16,340	2.89	0.71
	INTMAT—Interest in and enjoyment of mathematics	.90	16,347	2.32	0.84
	MATHEFF—Mathematics self-efficacy	.81	16,320	3.07	0.56
	SCMAT—Mathematics self concept	.91	16,299	2.55	0.83
	MEMOR—Memorization/rehearsal strategies	.68	16,255	2.54	0.64
	ELAB—Elaboration strategies	.74	16,266	2.35	0.61
	CSTRAT—Control strategies	.72	16,306	3.10	0.55
	COOPLRN—Preference for co-operative learning situations	.76	16,266	2.71	0.62
	COMPLRN—Preference for competitive learning situations	.84	16,292	2.56	0.69
Subsample 2	INSTMOT—Instrumental motivation to learn mathematics	.82	16,561	2.90	0.71
	INTMAT—Interest in and enjoyment of mathematics	.90	16,562	2.32	0.84
	MATHEFF—Mathematics self-efficacy	.81	16,524	3.07	0.56
	SCMAT—Mathematics self-concept	.91	16,506	2.54	0.82
	MEMOR—Memorization/rehearsal strategies	.68	16,454	2.55	0.64
	ELAB—Elaboration strategies	.73	16,478	2.36	0.60
	CSTRAT—Control strategies	.72	16,507	3.10	0.54
	COOPLRN—Preference for cooperative learning situations	.76	16,468	2.70	0.62
	COMPLRN—Preference for competitive learning situations	.84	16,501	2.56	0.69

Table A3. Multigroup Confirmatory Factor Analysis for Subsample 2.

Subsample 2		Model	MLR-χ	*df*	CFI	TLI	RMSEA	SRMS	*N*	Items	Factors
Group-specific baseline models	Students with German background (G)	1a	12817.18	909	.887	.877	.031	.066	13,741	45	9
		1b	7751.472	778	.929	.922	.026	.05	13,741	42	9
	Students with former-USSR background (f-U)	1c	2119.76	909	.878	.867	.03	.071	1,454	45	9
		1d	1479.731	778	.924	.916	.025	.058	1,454	42	9
	Students with Turkish background (T)	1e	1949.767	909	.854	.841	.031	.071	1,206	45	9
		1f	1372.514	778	.909	.899	.025	.06	1,206	42	9
Configural invariance	Parameters freely estimated across groups	2	9917.358	2,334	.927	.919	.024	.052	16,401	42	9
Metric invariance	Factor loadings constrained across groups	3	10006.45	2,400	.926	.921	.025	.052	16,401	42	9
Scalar invariance	Intercepts constrained across all groups	4a	10749.02	2,484	.92	.917	.025	.057	16,401	42	9
	Intercepts constrained across G and f-U students	4b	10275.82	2,442	.924	.92	.024	.053	16,401	42	9

(continued)

Table A3. (continued)

Subsample 2	Model	MLR-χ	*df*	CFI	TLI	RMSEA	SRMS	*N*	Items	Factors
Intercepts constrained across G and f-U students	4c	10166.23	2,429	.925	.92	.024	.053	16,401	42	9
Intercepts constrained across G and T students	4d	10591	2,442	.922	.917	.025	.056	16,401	42	9
Intercepts constrained across G and T students	4e	10426.96	2,429	.923	.918	.025	.055	16,401	42	9
Intercepts constrained across G and T students	4f	10374.9	2,425	.923	.918	.024	.054	16,401	42	9
Intercepts constrained across all groups	4g	10078.32	2,412	.926	.921	.024	.053	16,401	42	9

Note. G = German; f-U = former USSR; T = Turkish; MLR = maximum likelihood robust; CFI = comparative fit index; RMSEA = root mean square error of approximation; TLI = Tucker–Lewis index; SRMS = standardized root mean square residual. In Models 4c and 4e, only intercepts of MATHEFF, MEMOR, ELAB, CSTRAT, COOPLRN, and COMPLRN constrained across groups. In Model 4f, only intercepts of MEMOR, CSTRAT, and COOPLRN constrained across groups. Model 4g applies the constraints of Model 4c for students with a former USSR background and of Model 4f for students with a Turkish background.

Notes

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The international PISA student, teacher, and school databases are freely downloadable from http://pisa2000.acer.edu.au. The availability of the national data sets that may include country-specific extensions varies. For example, the German PISA-2003-E data set that was used in this study is made available for researchers through the Research Data Center (Forschungsdatenzentrum, FDZ) at the Institute for Educational Quality Improvement (Institut zur Qualitätsentwicklung im Bildungswesen, IQB, www.IQB.hu-berlin.de).
2. In most cases, researchers try to construct their items in a way that they have high loading on one factor but small and negligible loadings on other factors. Since in our model, each factor is specified as influencing only one item, we restrict the following discussion to cases where an item is a function of just one factor. However, theoretical models exist in which an item score is predicted by several factors.
3. Because of identification issues, some parameters have to be constrained. It is common practice to fix the factor loading of the first item of each factor to 1 (for a detailed discussion, see Brown, 2006; for a more critical discussion illustrating the potential limitations of this approach, see Raykov, Marcoulides, & Li, 2012).

References

Alba, R. D. (2008). Why we still need a theory of mainstream assimilation. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 48*, 37-56.

Avery, D. R., Tonidandel, S., Thomas, K. M., Johnson, C. D., & Mack, D. A. (2007). Assessing the multigroup ethnic identity measure for measurement equivalence across racial and ethnic groups. *Educational and Psychological Measurement*, *67*, 877-888.

Beckstead, J. W., Yang, C.-Y., & Lengacher, C. A. (2008). Assessing cross-cultural validity of scales: A methodological review and illustrative example. *International Journal of Nursing Studies*, *45*, 110-119.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). New York, NY: Cambridge University Press.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*, 161-186.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (2003). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Byrne, B. M., & Van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107-132.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*, 155-175.

Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, *81*, 78-96.

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

Christensen, G. S., & Segeritz, M. (2008). An international perspective on student achievement. In Bertelsmann Stiftung (Ed.), *Immigrant students can succeed* (pp. 11-33). Gütersloh, Germany: Bertelsmann Stiftung.

Fontaine, J. R. J. (2011). A fourfold conceptual framework for cultural and cross-cultural psychology: Relativism, construct universalism, repertoire universalism and absolutism. In F. J. R. Van De Vijver, A. Chasiotis, & S. M. Breugelmans (Eds.), *Fundamental questions in cross-cultural psychology* (pp. 165-189). Cambridge, England: Cambridge University Press.

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

Haller, W., Portes, A., & Lynch, S. M. (2011). Dreams fulfilled, dreams shattered: Determinants of segmented assimilation in the second generation. *Social Forces*, *89*, 733-762.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

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.

Johnson, T., Kulesa, P., Llc, I., Cho, Y. I., & Shavitt, S. (2005). The relationship between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*, 264-277.

Kao, G., & Tienda, M. (1995). Optimism and achievement: The educational performance of immigrant youth. *Social Science Quarterly*, *76*, 1-19.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: Wiley.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53-76.

Marcoulides, G. A., Emrich, C., & Marcoulides, L. D. (2008). Testing for multigroup invariance of the computer anxiety scale. *Educational and Psychological Measurement*, *68*, 325-334.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224-253.

Markus, H. R., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science*, *5*, 420-430.

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the big-two factor theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research*, *38*, 189-224.

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

Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin*, *34*, 543-552.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality*, *74*, 403-456.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592.

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

Mylonas, K. (2009). Statistical analysis techniques based on cross-cultural research methods: Cross-cultural paradigms and intra-country comparisons. *Psychology*, *16*, 185-204.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2005). *PISA 2003: Technical report*. Paris, France: Author.

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.

Peschar, J. L., Veenstra, R., Molenaar, I. W., Boomsma, A., Huisman, M., & van der Wal, M. (1999a). *Synthesis report: Self-regulated learning as a cross-curricular competency: The construction of instruments in 22 countries for the PISA main study 2000*. Groningen, Netherlands: University of Groningen.

Peschar, J. L., Veenstra, R., Molenaar, I. W., Boomsma, A., Huisman, M., & van der Wal, M. (1999b). *Technical report: Self-regulated learning as a cross-curricular competency: The construction of instruments in 22 countries for the PISA main study 2000*. Groningen, Netherlands: University of Groningen.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). San Diego, CA: Academic Press.

Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, *72*, 954-974.

Roche, C., & Kuperminc, G. P. (2012). Acculturative stress and school belonging among Latino youth. *Hispanic Journal of Behavioral Sciences*, *34*, 61-76.

Ryan, A. M., West, B. J., & Carr, J. Z. (2003). Effects of the terrorist attacks of 9/11/01 on employee attitudes. *Journal of Applied Psychology*, *88*, 647-659.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54-67.

Segeritz, M., Walter, O., & Stanat, P. (2010). Muster des schulischen Erfolgs von jugendlichen Migranten in Deutschland: Evidenz für segmentierte Assimilation? [Patterns of educational success of young immigrants in Germany: Evidence for segmented assimilation?]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*, 113-138.

Shavelson, R. J., & Byrne, B. M. (1987). Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal*, *24*, 365-385.

Silvia, P. J. (2006). *Exploring the psychology of interest*. New York, NY: Oxford University Press.

Simon, H. A., & Ericsson, K. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*, 178-186.

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, *137*, 421-442.

Stanat, P., & Christensen, G. S. (2006). *Where immigrant students succeed: A comparative review of performances and engagement in PISA 2003*. Paris, France: OECD.

Stanat, P., Segeritz, M., & Christensen, G. S. (2010). Schulbezogene Motivation und Aspiration von Schülerinnen und Schülern mit Migrationshintergrund. [Educational motivations and aspirations of students with an immigrant background.] In W. Bos, E. Klieme, & O. Köller (Eds.), *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert [Learning opportunities and development of competencies in the school context]* (pp. 31-57). Münster, Germany: Waxmann.

Takle, M. (2011). (Spät)Aussiedler: From Germans to immigrants. *Nationalism and Ethnic Politics*, *17*, 161-181.

Twenge, J. M., & Crocker, J. (2002). Race and self-esteem: Meta-analyses comparing Whites, Blacks, Hispanics, Asians, and American Indians and comment on Gray-Little and Hafdahl (2000). *Psychological Bulletin*, *128*, 371-408.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17-45). New York, NY: Cambridge University Press.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology*, *33*, 141-156.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39-47.

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. *Review of Research in Education*, *23*, 73-118.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1-26.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165-200.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press.