

# Chapter 5

## New Ways of Dealing with Lacking Measurement Invariance



Markus Sauerwein and Désirée Theis

### 5.1 Introduction

Over the past decade, policy-makers have become increasingly interested in studies, such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS), in which education systems of various countries are compared. Reforms in education are often based on or legitimated by results of such international studies, and governments may adopt educational practices common in countries that performed well in those studies in an attempt to improve their education system (Panayiotou et al., 2014).

Education can be analyzed at the student, classroom (or teacher), school, and (national) system levels (Creemers & Kyriakidēs, 2008, 2015). Decisions made at the system level (e.g. by policy-makers) affect all other levels. Information about, for example, student achievement or teaching quality in a given country can be compared to that in other countries and used to improve teaching quality. Thus, results of international studies in education, such as PISA, which provides information about students' academic achievement and teaching quality in more than 60 countries, are becoming increasingly interesting to policy makers and might affect classroom processes indirectly through reforms in education, and so on.

However, interpretation of the results of international studies may differ across cultures (Reynolds, 2006). Before a construct (of teaching quality), such as classroom management or disciplinary climate, can be compared across groups (e.g.

---

M. Sauerwein (✉)

Fließner University of Applied Sciences Düsseldorf, Düsseldorf, Germany  
e-mail: [sauerwein@dipf.de](mailto:sauerwein@dipf.de)

D. Theis

DIPF | Leibniz Institute for Research and Information in Education,  
Frankfurt am Main, Germany

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, [https://doi.org/10.1007/978-3-030-69345-9\\_5](https://doi.org/10.1007/978-3-030-69345-9_5)

63

countries), the structural stability of that construct needs to be investigated. Thus, measurement invariance (MI) analyses have to be conducted and scalar (factorial) invariance has to be established if mean level changes are to be compared across groups or over time (Borsboom, 2006; Chen, 2007, 2008; van de Schoot, Lugtig, & Hox, 2012).

Until now, MI has been neglected in many studies (e.g. Kyriakides, 2006b; OECD, 2012; Panayiotou et al., 2014; Soh, 2014), which could lead to a false interpretation of the implications of the results. In this paper, we analyze data of the PISA study to explore the effect of lacking MI in studies in which groups are compared. Moreover, we investigate whether lacking MI alone provides information about psychometric properties of the construct under investigation or if it also provides content-related information about the construct. We explore possible explanations for the missing MI by consulting third variables, which are very likely to be equivalent across countries.

### ***5.1.1 The Multi-Level Framework of the Education System***

Over the past decade, policy-makers and school administrators have shown an increasing interest in research findings concerning the association between teaching quality and student achievement (Pianta & Hamre, 2009a). Findings of studies, such as PISA, are used to justify and legitimize reforms in education (for a discussion about the influence of PISA findings on policy decisions, see Breakspear, 2012). Accordingly, one goal of studies, such as PISA (OECD, 2010; e.g. OECD Publishing, 2010, 2011) is to identify factors related to students' learning. Some of these factors can be influenced (indirectly) by changes in policy concerning, for example, the curriculum, resource allocation, or teaching quality (e.g. through teacher training or teacher education; Kyriakides, 2006a). The assumption that policy changes affect teaching quality, for example, is based on a multi-level framework of education systems.

The dynamic model of educational effectiveness (Creemers & Kyriakidēs, 2008, 2015; Creemers, Kyriakidēs, & Antoniou, 2013; Panayiotou et al., 2014) describes how system, school, and classroom levels interact. Scheerens (2016, p. 77) states that "within the framework of multi-level education systems, the school level should be seen from the perspective of creating, facilitating and stimulating conditions for effective instruction at the classroom level." Learning takes place primarily at the classroom level and is associated with teaching quality. At the school level, all stakeholders (teacher, parents, students, etc.) are expected to ensure that time in class is optimized and that teaching quality is improved (Creemers & Kyriakides, 2015). This way, the school level is expected to influence teaching quality (e.g. through regular evaluations at school). The school level, in turn, is influenced by the system/country level through education-related policy, systematic school and/or teacher evaluations, and teacher education (Creemers & Kyriakides, 2015). Hence, policies relevant not only at the classroom level but also at the school and/or country level can improve teaching quality.

### ***5.1.2 Context Matters: Comparing Educational Constructs in Different Contexts***

Since the beginning of the twenty-first century, policy-makers have attempted to transfer knowledge and ideas employed in one education system to another (Panayiotou et al., 2014). PISA provides information about students' academic achievement and teaching quality in more than 60 countries. The relation between students' academic achievement and teaching quality is worth being examined at the system level because low scores on achievement tests might correlate with poor teaching quality in a given country. Thus, when students perform poorly on achievement tests, policy-makers might be interested in comparing the teaching quality in their country to the teaching quality in other countries. Detailed knowledge about how students' academic achievement is promoted in various countries might help policy-makers develop appropriate teacher training programs.

As interest in international comparisons in education grows, researchers are becoming increasingly concerned that findings are too simplified and too easily transferred to different cultures (Reynolds, 2006). Comparison of education-related constructs in various subjects, grades, extracurricular activities, and countries requires MI across the different contexts. Hence, to legitimize comparisons of dimensions in different contexts, the dimensions must be stable across the given contexts. MI must be established for the construct under investigation in order to ensure this precondition.

### ***5.1.3 Teaching Quality***

Teaching quality often is framed according to the dynamic model of educational effectiveness (Creemers et al., 2013; Creemers & Kyriakidēs, 2008), the classroom assessment scoring system (CLASS) (Hamre & Pianta, 2010; Hamre, Pianta, Mashburn, & Downer, 2007; Pianta & Hamre, 2009a, 2009b), or the three dimensions of classroom process quality (Klieme, Pauli, & Reusser, 2009; Lipowsky et al., 2009; Rakoczy et al., 2007). These models, which show a considerable overlap (Decristan et al., 2015; Praetorius et al., 2018), refer to three essential generic dimensions of teaching quality. The first dimension can be described as classroom management (see also Kounin, 1970) or disciplinary climate. This dimension is closely related to the concept of time on task. It is postulated that clear structures and rules can help students to focus on lessons and to complete tasks (Doyle, 1984, 2006; Evertson & Weinstein, 2006; Kounin, 1970; Oliver, Wehby, & Daniel, 2011). Several studies and meta-analyses have shown a positive correlation between classroom management and students' learning (Hattie, 2009; Kyriakides, Christoforou, & Charalambous, 2013; Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993). The second dimension is cognitive activation or instructional support and refers to (constructivist) learning theories (Fauth, Decristan, Rieser, Klieme, & Büttner,

2014; Klieme et al., 2009; e.g. Lipowsky et al., 2009; Mayer, 2002). The third dimension is commonly referred to as supportive climate, emotional support (e.g. Klieme et al., 2009; Klieme & Rakoczy, 2008), or students' motivation (e.g. Kunter & Trautwein, 2013) and is derived from motivation theories, self-determination theory, in particular (Deci & Ryan, 1985; Ryan & Deci, 2002). In this chapter, we focus on disciplinary climate as a subdimension of classroom management – one central dimension of teaching quality, which is assessed in PISA.

### 5.1.4 *Measurement Invariance Analyses*

Generally, MI analyses are conducted to determine the psychometric properties of scales and constructs. MI of the construct under investigation across two or more groups or assessment points must be established when (mean) scores of scales, or the influence of a variable on another, are compared because such analyses postulate that the scale measures the same construct in all groups over a certain period of time. If MI is not established, the scale will not measure the same construct in all groups. The results of such comparisons in which MI is not established might be biased and cannot be interpreted as originally intended (Borsboom, 2006; Chen, 2007, 2008; van de Schoot et al., 2012).

MI needs to be distinguished from measurement bias: While bias refers to differences between the estimated parameter and the true parameter, MI refers to comparability across groups (Sass, 2011). Generally, three levels of MI can be differentiated. The most basic level of MI is configural invariance, which is established when items are associated with the same latent construct in different groups or across assessment points. If configural invariance is established, the scale will measure similar but not equal constructs across groups/assessment points. In this case, comparisons of correlations between the scale and other variables in different groups are legitimate. Effect sizes of these correlations, however, should not be interpreted and compared. If configural invariance was not established, scores on the scale under investigation should not be compared across groups or assessment points. The second level of MI is called metric invariance, which is established when factor loadings are equal across groups or assessment points. Value changes in an item for one unit lead to equal changes in the latent construct for all groups. This level of MI allows comparison of associations (and effect sizes) between latent scales and variables across groups or assessment points (Vandenberg & Lance, 2000; Vieluf, Leon, & Carstens, 2010). The third level of MI is scalar invariance, which is established when factor loadings and intercepts of the items representing the latent construct are equal across groups or assessment points. Therefore, the scales share the same intercept. Thus, all groups under investigation have the same starting point, and mean scores can be compared (Chen, 2008; Vandenberg & Lance, 2000).

Recent studies show that the necessary level of measurement invariance for cross-cultural comparisons often is not given (e.g. Vieluf et al., 2010). Moreover,

some studies do not even control for or report MI. Luyten et al. (2005) found that the interactions between socio-economic status (SES) and teaching quality differ across countries, but the authors do not report whether the necessary level of MI (here at least metric MI) for cross-cultural comparisons was established. Similarly, Panayiotou et al. (2014) test the dynamic model of educational effectiveness in different countries and compare the influence of several factors on student achievement, but do not investigate the level of MI for their construct among the different countries (only within the countries) (see also Kyriakides, 2006b and Soh, 2014).

### 5.1.5 Research Objectives

As mentioned above, results of studies investigating differences in teaching quality across countries are of great interest to policy-makers. Information provided by such studies affects decisions that are made at the system level, which, in turn, affect processes at the classroom level. However, in order to compare certain constructs across groups or over time, invariance among the scales under investigation must be established, which, until now, has not necessarily been the case. The objectives of the present chapter are to

- show how neglecting MI of dimensions under investigation affects results of studies, in which mean levels among groups or assessment points are compared;
- compare the mean score of disciplinary climate among countries;
- investigate whether constructs can be compared even if a certain level of MI is not established; and
- find variables, which could explain the lack of MI among countries.

## 5.2 Method

### 5.2.1 Study

We analyzed data from PISA 2009; PISA is a triennial international comparative study of student learning outcomes in reading, mathematics, and science. The focus in PISA 2009 was reading comprehension, which we used as the outcome variable. The reading test in PISA is set at a mean ( $M$ ) of 500 points and a standard deviation ( $SD$ ) of 100 points. The study originally was developed as an instrument for OECD countries; now, it is used in more than 65 countries. The study is designed to monitor outcomes over time and provides insights into the factors that may account for differences in students' academic achievement within and among countries (OECD, 2011, 2012).

Students complete a questionnaire assessing, for example, classroom management (measured as disciplinary climate) in the native language lesson (OECD,

2012). Table 5.1 shows the items assessed with this scale (1 = *strongly disagree* – 4 = *strongly agree*) and sample size, means, and the standard deviation of students from Chile, Finland, Germany, and Korea, who participated in PISA 2009. We refer to these countries because they are typical proxies for region-specific educational systems.<sup>1</sup> Furthermore, we use class size as the measurement equivalent variable to explain lacking MI among the countries. For mean and standard deviation of the variable *class size*, see Table 5.2.

**Table 5.1** Descriptive statistics of the scale used to assess disciplinary climate in PISA

		Students don't listen to what the teacher says	There is no noise or disorder	The teacher has to wait a long time for students to quiet down	Students cannot work well	Students don't start working for a long time after lessons begin
Chile	<i>M</i>	2.14	2.34	2.22	1.84	2.12
	<i>N</i>	5550	5554	5551	5554	5555
	<i>S.D.</i>	.743	.812	.907	.805	.879
Finland	<i>M</i>	2.40	2.49	2.27	1.94	2.19
	<i>N</i>	5770	5770	5769	5765	5767
	<i>S.D.</i>	.764	.824	.848	.783	.866
Germany	<i>M</i>	1.90	1.86	2.02	1.88	1.84
	<i>N</i>	4420	4430	4424	4390	4417
	<i>S.D.</i>	.780	.830	.871	.838	.888
Korea	<i>M</i>	1.80	2.11	1.72	1.63	1.71
	<i>N</i>	4966	4962	4962	4961	4964
	<i>S.D.</i>	.631	.681	.714	.697	.729
All	<i>M</i>	2.08	2.23	2.07	1.83	1.98
	<i>N</i>	20,706	20,716	20,706	20,670	20,703
	<i>S.D.</i>	.768	.824	.867	.790	.866

*M* Mean, *S.D.* Standard deviation, *N* Number of students

**Table 5.2** Class size

	<i>N</i>	<i>M</i>	<i>SD</i>
Chile	5189	36.16	7.56
Finland	5643	18.77	4.13
Germany	4200	24.66	5.17
Korea	4986	35.98	5.07
All	20,018	28.80	9.51

*M* Mean, *S.D.* Standard deviation, *N* Number of students

<sup>1</sup>Chile represents a South-American system with highly improved rates in PISA tests in the last decades; Germany is well-known for its highly structured education system and is, besides Finland, used as an example for a European system. Korea is a proxy for an Eastern-Asian system with a strong focus on performance and good PISA results. Finland is used as an example for a Scandinavian system, and students are also performing very well in PISA studies.

### 5.2.2 Data Analyses

Below is a step-by-step explanation of how we compared the scales of the different countries.

#### 1. Comparison of mean levels and associations between disciplinary climate and reading

First, we performed an analysis of variance (ANOVA) to compare mean levels. This allowed us to determine whether there were significant differences in disciplinary climate among the countries. Cohen's  $d$  was used to indicate the magnitude of the differences among the countries. Values between .2 and .5 indicated small effect sizes; values between .5 and .8 indicated moderate effect sizes. Higher values (>.8) indicated large effect sizes (Cohen, 1988). Second, we computed regression analyses to identify the association between reading score and disciplinary climate. Including this step before the MI analyses shows how false conclusions can be drawn, if mean levels are compared although MI is lacking. Normally, MI has to be established before mean level scores and effect sizes are compared. However, we turned the normal procedure around in favour of our research objectives.

#### 2. MI analyses and explaining lack of MI

We conducted MI analyses to test the structural stability of the scales used in the context of PISA. A model with parameter constraints was tested against a less restricted model (e.g. metric vs. configural invariance). To determine the level of MI, we compared the fit indices of the models. In line with the literature at hand, we used the comparative fit index (CFI), and the root mean square error of approximation (RMSEA) to test, which model fit the data best (Chen, 2007; Desa, 2014; Sass, 2011; Sass, Schmitt, & Marsh, 2014; Vandenberg & Lance, 2000; Vieluf et al., 2010). A model was accepted, if the fit indices obtained the following scores: CFI > .90, RMSEA < .08 (Hu & Bentler, 1999). In line with results of simulation studies, Chen (2007) recommends that the next higher level of MI be revised, if the CFI decreases by  $\geq -.01$  and/or the RMSEA decreases by  $\geq .015$ . However, Chen (2007, p. 502) states that "[...] these criteria should be used with caution, because testing measurement invariance is a very complex issue." Another way to determine the level of MI is to conduct a chi-square test; however, the results of these tests should be interpreted with caution as they are influenced by sample size. Thus, models designed on the basis of a large sample size could be rejected even if they fit the data well (van de Schoot et al., 2012; Vandenberg & Lance, 2000). The sample studied in PISA is quite large. Thus, we did not conduct chi-square tests. We investigated whether scales or at least single items could be compared among countries. Therefore, we performed the analyses as follows:

- First, we determined the level of MI across all four countries we refer to in our paper (Korea, Finland, Germany, and Chile).
- Second, we determined the level of MI when countries were compared.

- Third, we examined the factor loadings ( $\lambda$ ) of the items and investigated whether single items had the same or different (content-related) meaning for the latent construct. To decide, which items had different meanings in different countries, we used the MODINDICES function in MPlus 7.1 (see Muthén & Muthén, 1998–2012). The MODINDICES function provides information about fixed items (between groups) and the expected improvement of model fit if a certain item is freely estimated. Items, which could be fixed between groups, seemed to have the same relevance or meaning for the latent construct in different countries.
- Fourth, we investigated whether single items were comparable. Therefore, we established partial MIs: Some of the factor loadings and/or intercepts among groups were allowed to be estimated freely, while others remained constant (van de Schoot et al., 2013). To decide, which items should be estimated freely, we used again the MODINDICES function in Mplus (Muthén & Muthén, 1998–2012). We allowed factor loadings or intercepts among groups of some items to be estimated freely until the model showed an acceptable fit. This approach allowed us to find items, which were comparable among countries.
- Finally, we tried to identify the reason for possible lacks in MI. We considered variables, which were measurement-invariant by definition among countries. For the purpose of this study, we used the variable class size (see Table 5.2) because a student is a student in every country and therefore comparable across countries.

## 5.3 Results

### 5.3.1 *Research Aim No. 1: How Neglecting MI Could Lead to False Interpretations of Results*

Table 5.3 shows the mean levels of the different countries on the scale used to assess disciplinary climate. Without taking MI into account, these results indicate that the highest level of disciplinary climate was reported in Korea. As all differences among the countries are significant ( $p < .01$ ), we also calculated Cohen's  $d$ . Our results indicate that there are moderate differences in terms of the mean scores of disciplinary climate between Chile and Korea, Finland and Germany, and Finland and Korea. Moreover, our results show that students in Finland and Korea achieved the highest scores in reading competence (Korea: 539; Finland: 536) (OECD, 2011), but disciplinary climate in both countries differed significantly (Table 5.3). Therefore, we also computed regression analyses to explain the relation between disciplinary climate and reading competence.

As shown in Table 5.4, we found differences in the predictive value of disciplinary climate/classroom management among the countries; in Finland, this effect was very small. Policy-makers in Chile might conclude from these findings that the concept of disciplinary climate in Korea should be adopted in Chile. However, before such conclusions can be drawn, it needs to be tested whether disciplinary climate



**Table 5.3** Cohen’s *d* and scores on the reading test

	N	Disciplinary climate Mean	Cohen’s <i>d</i> (differences among the countries)				Reading score – mean
			Chile	Finland	Germany	Korea	
Chile	5567	2.13	–	–0.19	0.35	0.56	449
Finland	5774	2.26	0.19	–	0.53	0.76	536
Germany	4443	1.90	–0.35	–0.53	–	0.17	497
Korea	4972	1.79	–0.56	–0.76	–0.17	–	539

Note: *N* number of students

**Table 5.4** Effect of disciplinary climate on reading competences

	<i>B</i>	<i>R</i> <sup>2</sup>
Chile	–14.20	0.03
Finland	–6.33	0.01
Germany	–19.48	0.04
Korea	–14.49	0.04

*B* unstandardized effect of disciplinary climate on Reading Competences (Note: PISA Reading Competence Test has a mean of 500 and a standard deviation of 100)

**Table 5.5** MI analyses across countries

	Configural invariance	Metric invariance
CFI	.991	.906
RMSEA	.041	.099

*CFI* Comparative Fit Index, *RMSEA* Root Mean Square Error of Approximation

has the same meaning in the countries (i.e. Chile and Korea). Therefore, we investigated whether this scale was stable across the different countries, and if mean levels were, thus, comparable.

### 5.3.2 *Research Aim No. 2: Investigating the Stability of the Scale Used to Assess Disciplinary Climate Across Countries and Comparing Countries Even if MI Is Missing*

First, we determined the level of MI across all four countries. Table 5.5 shows that configural MI was established because there was a meaningful decrease in model fit when we tested the model with greater constraints (metric invariance). This result indicates that mean scores of the latent construct of disciplinary climate cannot be interpreted. The same holds true for the association between this construct and other variables. Thus, it is not legitimate to conclude that the effect of disciplinary climate

on reading competence in Germany is larger than in Finland. In all countries, a similar but not the same construct was measured and solely comparisons of the direction of correlations were legitimate. Hence, one might conclude that there was a positive correlation between students' achievement and disciplinary climate in all countries.

Second, we examined the comparability of countries and ran MI analyses separately for each possible comparison option among the four countries. Table 5.6 illustrates that a comparison of the mean scores between Finland and Chile was legitimate. Here, a better disciplinary climate was reported for Chile ( $M = 2.13$ ) than for Finland ( $M = 2.26$ ). A comparison of the effects of disciplinary climate between Finland and Korea as well as between Chile and Korea was legitimate. In the last case, the model fit (i.e. the CFI and RMSEA) decreased by more than .01. Nonetheless, the fit was acceptable and a comparison might have been legitimate. Thus, here we were able to compare the strength of the relation between disciplinary climate and student achievement.

We found a stronger relation between disciplinary climate and reading competency in Korea than in Finland. In Korea and Chile, the strength of the relation was comparable (see Table 5.4). Comparisons between the other countries were not possible because the necessary level of MI was not established.

Third, we investigated whether the factor loadings of single items in different countries might be interpreted. Table 5.7 shows the factor loadings of the single items. Using the MODINDICES function in MPlus, we were able to conclude from our findings that, for example, items 1 and 2 caused meaningful decreases in the

**Table 5.6** Investigating MI among countries

	Configural MI	Metric MI	Scalar MI
Chile – Korea			
CFI	0.990	0.974	.934
RMSEA	0.041	0.054	.075
Chile – Germany			
CFI	0.996	0.927	–
RMSEA	0.028	0.093	–
Germany – Finland			
CFI	0.991	0.904	–
RMSEA	0.042	0.111	–
Chile – Finland			
CFI	.988	.986	.976
RMSEA	.054	.048	.054
Finland – Korea			
CFI	0.985	0.977	0.927
RMSEA	0.053	0.055	0.084
Korea – Germany			
CFI	.994	.880	
RMSEA	.029	.112	

CFI Comparative Fit Index, RMSEA Root Mean Square Error of Approximation, MI Measurement Invariance

**Table 5.7** Comparison of factor loadings

	$\lambda$ factor	<i>S.E.</i>
Chile		
Item 1 – Students don't listen to what the teacher says	0.798	0.015
Item 2 – There is no noise or disorder	0.848	0.011
Item 3 – The teacher has to wait a long time for students to quiet down	0.838	0.010
Item 4 – Students cannot work well	0.815	0.013
Item 5 – Students don't start working for a long time after lessons begin	0.838	0.010
Finland		
Item 1 – Students don't listen to what the teacher says	0.830	0.011
Item 2 – There is no noise or disorder	0.873	0.008
Item 3 – The teacher has to wait a long time for students to quiet down	0.873	0.010
Item 4 – Students cannot work well	0.777	0.017
Item 5 – Students don't start working for a long time after lessons begin	0.825	0.012
Germany		
Item 1 – Students don't listen to what the teacher says	0.914	0.007
Item 2 – There is no noise or disorder	0.955	0.005
Item 3 – The teacher has to wait a long time for students to quiet down	0.944	0.005
Item 4 – Students cannot work well	0.894	0.009
Item 5 – Students don't start working for a long time after lessons begin	0.924	0.006
Korea		
Item 1 – Students don't listen to what the teacher says	0.740	0.028
Item 2 – There is no noise or disorder	0.716	0.027
Item 3 – The teacher has to wait a long time for students to quiet down	0.726	0.025
Item 4 – Students cannot work well	0.858	0.014
Item 5 – Students don't start working for a long time after lessons begin	0.845	0.013

$\lambda$  factor Factor Loading, *S.E.* Standard Error

model fit (the respective values are not reported on in the table) when Chile and Germany were compared. In the case of Finland and Germany, items 1 and 4 led to a decrease in the model fit. Moreover, items 2 and 3 differed from each other when Korea and Finland were compared. However, here no meaningful decrease in the model fit was found.

Taking Germany and Chile as examples, the MODINDICES in MPlus indicated that fixing the factor loadings of items 1 and 2 led to a decline in model fit. Furthermore, it can be seen in Table 5.6 that the factor loadings for these items differed. To avoid a decline in model fit, we calculated partial metric MI (see van de Schoot et al., 2013). Here, the factor loadings of items 1 and 2 were estimated freely (CFI: .94; RMSEA: .09). Next, we used the MODINCES function again to decide whether more items needed to be estimated freely. However, the analyses produced no model with a satisfying model fit. Thus, mean scores of the scale to assess disciplinary climate in Germany and Chile could not be compared (even if we had merely fixed the factor loading of one item). In the same way, we freely estimated factor loadings between Chile and Korea. Here, the analysis would produce a satisfying

model fit, if we fixed the factor loading of item 4 only (CFI: .99; RMSEA: .04). Hence, a comparison of Chile and Korea for this item (“Students cannot work well”) was justified. Accordingly, we conducted a regression analysis while testing the predictive value of this item in terms of the reading achievement of students in Korea and those in Chile. Results of this analysis indicate that the item had greater predictive value in terms of the Korean students’ achievement in reading than in the reading achievement of the Chilean students (Korea:  $B = -16.05$ ; Chile:  $B = -13.93$ ). Even when the intercept of item four was fixed between Korea and Chile, no meaningful decrease in model fit was found (CFI: .98; RMSEA: .05). Thus, mean scores of this item could be compared between Korea and Chile (Chile:  $M = 1.84$ ; Korea:  $M = 1.63$ ;  $p < .01$  | Cohen’s  $d = .28$ ).

Our findings indicate that merely fixing this item led to an acceptable model fit (the factor loadings of all other items were estimated freely). Thus, Chile and Korea can be compared in terms of this single item only even when comparison of single items is seen as critical. Nonetheless, results of the regression analyses indicate that comparing the predictive value of a single item can provide meaningful results. If no comparisons were allowed, however, an interpretation of the different meanings of the items in cultural contexts could be worthwhile. For example, if we wanted to compare Germany and Chile, results of the analysis would indicate that no comparisons are allowed. However, we could say that item 1 (“Students don’t listen to what the teacher says”) is more relevant for the latent construct of disciplinary climate in Germany than in Chile (by comparing factor loadings), and this could be an interesting result on its own.

### ***5.3.3 Research Aim No. 3: Explaining Missing MI by Using Other Variables, Which Are Considered to Have the Same Meaning in Different Countries***

Since the meaning of disciplinary climate varied somewhat across the countries under investigation, we searched for possible cultural explanations for the differences in meaning. The challenge here was to find a third variable that definitely had the same meaning in all countries, in other words, a variable, which was measurement-invariant. Thus, if we tried to explain the cultural differences in the meaning of disciplinary climate across the countries by another variable, this variable ought to be culture-invariant so that it can be used as an anchor. One variable that was invariant across the countries under investigation was the number of students in class. This item has the same zero point (=intercept) and the same factor loadings in every country, because a student is counted as one student everywhere and therefore leads to the same decrease of the scale class size. Furthermore, research and practitioners might suggest that classroom size and disciplinary climate are correlated. Thus, we used the number of students in class as an anchor

when trying to explain the cultural differences in the concept of disciplinary climate. We conducted several regression analyses: We used the entire scale as a dependent variable and five single items related to disciplinary climate as dependent variables. In all models, the number of students was used as the independent variable. We conducted these analyses separately for Chile, Finland, Korea, and Germany.

In Chile and Finland, the number of students in class predicted disciplinary climate (see Table 5.8). In these countries, disciplinary climate became more problematic as the number of students in class increased. We found the opposite effect in Korea: A large number of students in class correlated positively with disciplinary climate. In Finland and Chile, the number of students in class also correlated with items 2, 3, and 5. In Korea, the opposite effect was found when item 2 was used as the outcome variable. For Germany, we found no effects.

In summary, our results indicate that the number of students in class can be used as a variable to explain why disciplinary climate has the same meaning (scalar) in Chile and Finland and why, thus, mean levels are comparable in these countries. In these countries, disciplinary climate is associated with the same invariant third variable, and this might – but not must – be a reason why we find scalar MI between Chile and Finland. Furthermore, we found that comparisons of mean scores or correlations between disciplinary climate and other variables (e.g. reading comprehension) were not legitimate between Germany and other countries. Here, class size had no effect on disciplinary climate, which supports our interpretation described above. In Korea, the effects of number of students in class were inversed to Finland and Chile but still had predictive value. This might be the reason why disciplinary climate had a similar meaning in these countries (metric MI) but not the same meaning, which allows mean score comparisons; mean level comparisons were not allowed. However, we can compare the relation between disciplinary climate and reading competencies in Korea with that in Chile, and Finland.

**Table 5.8** Regression analysis: independent variable = number of students in class; dependent variable = scale of disciplinary climate as well as the single items of scale separately

	Chile <i>B</i>	Finland <i>B</i>	Korea <i>B</i>	Germany <i>B</i>
Disciplinary Climate	.030*	.074***	-.053***	-.012
Item 1 – Students don't listen to what the teacher says	.019	.103***	-.018	.020
Item 2 – There is no noise or disorder	.030*	.074***	-.053***	-.012
Item 3 – The teacher has to wait a long time for students to quiet down	.042**	.088***	.019	.012
Item 4 – Students cannot work well	.007	.012	-.024	.002
Item 5 – Students don't start working for a long time after lessons begin	.027*	.041***	-.022	.027

Note: \* =  $p < .05$ , \*\* =  $p < .01$ , \*\*\* =  $p < .001$

## 5.4 Discussion

Our results underline the importance of MI analyses in international comparative educational studies. Analyses based on PISA 2009 data show that results of such studies might be biased or misinterpreted, if MI was not tested before any further analyses are conducted. However, our findings also suggest that more detailed analyses would be worthwhile.

If MI was ignored, our findings indicated that students in Finland and Korea achieve high scores in terms of reading achievement while the mean level of disciplinary climate differed significantly between these countries. Moreover, the predictive value of disciplinary climate for the students' reading achievement differed significantly between these countries as well. Especially in Finland, the effect of disciplinary climate on reading achievement was rather low. The finding that classroom management (disciplinary climate) was an important predictor for students' learning is in line with findings from earlier studies (Carroll, 1963; Seidel & Shavelson, 2007). Such findings might be particularly valuable to policy-makers. For example, policy-makers in Germany might conclude that in good education systems, like the one in Finland, disciplinary climate is not relevant for student achievement. As a result, disciplinary climate might not be included as an indicator of teaching quality in schools or teacher evaluations anymore. However, these findings need to be treated with caution as they stem from analyses that are not legitimate from a methodological point of view. Analyses and interpretations, as they were described in this section, postulate that the constructs under investigation have the same meaning across groups. MI analyses, however, indicate that only configural MI was established in the scales we used; thus, mean levels in the different countries cannot be compared. Nonetheless, we recommend further analyses to be conducted in which findings from different countries will be compared. Additionally, our findings indicate that analyzing levels of MI based on single items can be worthwhile: In Chile – for the factor disciplinary climate – it is important to be quiet during lessons (item 2), and that teachers do not have to wait too long until lessons can start (item 3). If Germany and Chile were compared, it seemed that in Germany, the first item (“Student’s don’t listen to the teacher”) as well as the second item (“There is no noise or disorder”) were more relevant for the disciplinary climate. Comparing Finland and Germany showed that in Finland, item 1 (“Students don’t listen to what the teacher says”) and item 4 (“Can’t work well”) were not as meaningful as they were in Germany. The interpretation of factor loadings as a result on its own seems to be uncommon. However, this idea is similar to interpretations of differential item functioning (DIF) in the context of test construction and scaling (Klieme & Baumert, 2001; see also Greiff & Scherer, 2018). One possible explanation for differences in factor loadings could be that students in different countries/cultures have a different system of relevance for disciplinary climate, and therefore the meaning of disciplinary climate differs among countries/cultures. Teaching and behaviour during class are liable to cultural contexts. This is also underlined by different factor loadings.

If a construct compared between two groups does not meet the standards of MI, the construct conceptually conveys different meanings in these groups (Chen, 2008). Creemers and Kyriakides (2009), for example, report that the development of a school policy for teaching and evaluation has stronger effects in schools where the quality of teaching at the classroom level is low. However, this conclusion could be drawn only if a necessary level of MI was established, otherwise the conclusion drawn may be wrong. If research on school improvement and school effectiveness aimed to compare models in different countries – such as the dynamic model of educational effectiveness – the level of MI should be investigated and proven as a precondition of further analyses. A good example of how to determine and deal with MI in international studies has been described in a very detailed technical report of the TALIS study (OECD, 2014; Vieluf et al., 2010). Moreover, even if MI is missing for the entire scale, it is possible to identify single countries or items for comparison. As a preliminary step, not a multi-group CFA should be conducted with all countries in one model, but rather single countries should be selected for comparison. This might help researchers identify several countries for comparison. If scalar invariance is not given in the countries under investigation, it would be possible to identify single items that can be compared in a next step.

The analyses presented in this paper show that missing MI is not a reason for desisting from comparisons (between pedagogical contexts or cultures). Our findings indicate that the meaning of disciplinary climate differs among cultural contexts. In our opinion, this result should also be reported as a result of its own (see also Greiff & Scherer, 2018, for that issue). Given the fact that research in education is used as a tool to legitimate policy actions and that results are transferred from one cultural context to another, reporting missing MI appears to be especially important (Martens & Niemann, 2013; Panayiotou et al., 2014; Reynolds, 2006). Even if schools within a country were compared, MI should be tested because all schools differ from one another and might have their own school culture. Therefore, conclusions that the development of a school policy for teaching and external evaluation have been found to be more influential in schools where the quality of teaching at the classroom level is low (Creemers & Kyriakides, 2009) should be treated with caution.

Furthermore, qualitative methods (e.g. documentary methods, such as comparative analyses of different milieus, fields, cultural experiences, etc.; Bohnsack, 1991) refer to different systems of relevance people have, due to different structures of everyday life. The aim of this method is not to compare certain manifestations or means but rather to explain differences. This methodological background can be used to interpret the result of missing MI. In the case of lessons, we can assume that students have different systems of relevance when they are rating classroom management or disciplinary climate. In other words, students do not refer to the same standards when they rate lessons. Thus, we have good reasons to interpret missing MI as an important result. Theoretically, this reasoning is also in line with Lewin's field theory (Lewin, 1964). Person, context, and environment influence and depend on each other. Hence, teaching quality is nested in its cultural and pedagogical context. "Teachers' work does not exist in a vacuum but is embedded in social, cultural,

and organizational contexts” (Samuelsson & Lindblad, 2015, p. 169). A high-quality teacher in India does not allow questioning by students whereas in classes in the United States of America, the opposite is true (Berliner, 2005). Differences in factor loadings and intercepts could be seen as an expression of the cultural and institutional varieties, which should be considered more in international comparative studies. Furthermore, new possibilities may present themselves to identify what cultures display similar facets of teaching, schools, and the education system and therefore what characteristics thereof could be transferred to other education systems.

## 5.5 Conclusion

This paper presents one of the first attempts to interpret (lacking) MI not only from a methodological point of view but also in terms of content. Chen (2008) explains missing MI for the construct *self-esteem* between China and the USA. Our results indicate that the lack of MI can be seen as a result as well. Nevertheless, we propose further analyses that might investigate ways to compare at least parts of constructs. In summary, our approach to interpreting MI is in line with those of many researchers investigating school improvement and school development, who emphasize the local context of schools and stress the importance of international comparisons (Hallinger, 2003; Harris, Adams, Jones, & Muniandy, 2015; e.g. Reynolds, 2006). The analyses presented here make it possible to identify comparable single cross-cultural items.

## References

- Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, 56(3), 205–213. <https://doi.org/10.1177/0022487105275904>
- Bohnsack, R. (1991). *Rekonstruktive Sozialforschung. Einführung in Methodologie und Praxis qualitativer Forschung*. Opladen: Leske + Budrich.
- Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, 44(Suppl 3), S176–S181. <https://doi.org/10.1097/01.mlr.0000245143.08679.cc>
- Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance* (OECD Publishing, Ed.) (OECD Education Working Papers no. 71).
- Carroll, J. (1963). A model of school learning. *Teachers College Record*, 64, 723–733.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5), 1005–1018. <https://doi.org/10.1037/a0013193>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale: L. Erlbaum Associates.
- Creemers, B., & Kyriakides, L. (2009). Situational effects of the school factors included in the dynamic model of educational effectiveness. *South African Journal of Education*, 29, 293–315.



- Creemers, B., & Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, 26(1), 102–119. <https://doi.org/10.1080/09243453.2013.869233>
- Creemers, B. P. M., & Kyriakides, L. (2008). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. In *Contexts of learning*. London, UK/New York, NY: Routledge.
- Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). *Teacher professional development for improving quality of teaching*. Dordrecht, The Netherlands/New York, NY: Springer.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior: Perspectives in social psychology*. New York, NY: Plenum.
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Buttner, G., Fauth, B., ... Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, 52(6), 1133–1159. <https://doi.org/10.3102/0002831215596412>
- Desa, D. (2014). *Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses* (OECD education working papers, no. 103). Retrieved from OECD website: <https://doi.org/10.1787/5jz2kbbvlb7k-en>
- Doyle, W. (1984). How order is achieved in classrooms: An interim report. *Journal of Curriculum Studies*, 16(3), 259–277. <https://doi.org/10.1080/0022027840160305>
- Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management. Research, practice, and contemporary issues* (pp. 97–125). Mahwah, NJ: Lawrence Erlbaum Associates.
- Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>
- Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? *European Journal of Psychological Assessment*, 34(3), 141–144.
- Hallinger, P. (2003). Leading Educational Change: reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, 33(3), 329–351.
- Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25–41). New York, NY: Routledge.
- Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms* (Foundation for Child Development, Ed.).
- Harris, A., Adams, D., Jones, M. S., & Muniandy, V. (2015). System effectiveness and improvement: The importance of theory and context. *School Effectiveness and School Improvement*, 26(1), 1–3. <https://doi.org/10.1080/09243453.2014.987980>
- Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement*. London, UK: Routledge.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, 16(3), 385–402. <https://doi.org/10.1007/BF03173189>
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster, Germany: Waxmann.

- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcomeorientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237. Accessed 29 Dec 2012. [https://www.pedocs.de/frontdoor.php?source\\_opus=4348](https://www.pedocs.de/frontdoor.php?source_opus=4348)
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York, NY: Holt Rinehart & Winston.
- Kunter, M., & Trautwein, U. (Eds.). (2013). *Standard Wissen Lehramt: Vol. 3895. Psychologie des Unterrichts*. Paderborn, Germany: Schöningh.
- Kyriakides, L. (2006a). Introduction international studies on educational effectiveness. *Educational Research and Evaluation*, 12(6), 489–497. <https://doi.org/10.1080/13803610600873960>
- Kyriakides, L. (2006b). Using international comparative studies to develop the theoretical framework of educational effectiveness research: A secondary analysis of TIMSS 1999 data. *Educational Research and Evaluation*, 12(6), 513–534. <https://doi.org/10.1080/13803610600873986>
- Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, 36, 143–152.
- Lewin, K. (1964). *Field theory in social science*. New York, NY: Harper & Brothers.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. <https://doi.org/10.1016/j.learninstruc.2008.11.001>
- Luyten, J. W., Scheerens, J., Visscher, A. J., Maslowski, R., Witziers, B. U., & Steen, R. (2005). *School Factors related to quality and equity. Results from PISA 2000* (OECD, Ed.).
- Martens, K., & Niemann, D. (2013). When do numbers count?: The differential impact of the PISA rating and ranking on education policy in Germany and the US. *German Politics*, 22(3), 314–332. <https://doi.org/10.1080/09644008.2013.794455>
- Mayer, R. E. (2002). Understanding conceptual change: A commentary. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change. Issues in theory and practice* (pp. 101–111). Dordrecht, The Netherlands: Kluwer Academic.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus statistical analysis with latent variables: User's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- OECD. (2010). *PISA 2009 results*. Paris, France: OECD.
- OECD. (2011). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science (volume I) (1. Aufl.)*. PISA 2009 results. s.l. Paris, France: OECD.
- OECD. (2012). *PISA 2009 technical report*. Paris, France: OECD.
- OECD. (2014). *TALIS 2013 technical report*.
- Oliver, R., Wehby, J., & Daniel, J. (2011). *Teacher classroom management practices: Effects on disruptive or aggressive student behavior*. Campbell Systematic Reviews 2011.4.
- Panayiotou, A., Kyriakides, L., Creemers, B. P. M., McMahon, L., Vanlaar, G., Pfeifer, M., ... Bren, M. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability*, 26(1), 73–93. <https://doi.org/10.1007/s11092-013-9182-x>
- Pianta, R. C., & Hamre, B. K. (2009a). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>
- Pianta, R. C., & Hamre, B. K. (2009b). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, 2009(121), 33–46. <https://doi.org/10.1002/yd.295>
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality. The German framework of Three Basic Dimensions. *ZDM*, 50, (3, 407–426). <https://doi.org/10.1007/s11858-018-0918-4>
- Publishing, O. (2010). *PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices (Volume III) (1. Aufl.)*. PISA 2009 Results. s.l. Paris, France: OECD.

- Publishing, O. (2011). *PISA quality time for students: Learning in and out of school (1. Aufl.)*. PISA. s.l. Paris, France: OECD.
- Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007). Structure as a quality feature in mathematics instruction: Cognitive and motivational effects of a structured organisation of the learning environment vs a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG priority programme* (pp. 101–120). Münster, Germany/New York, NY/München, Germany/Berlin, Germany: Waxmann.
- Reynolds, D. (2006). World Class Schools: Some methodological and substantive findings and implications of the International School Effectiveness Research Project (ISERP). *Educational Research and Evaluation*, 12(6), 535–560. <https://doi.org/10.1080/13803610600874026>
- Ryan, R., & Deci, E. (2002). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–33). Rochester, NY: University of Rochester Press.
- Samuelsson, K., & Lindblad, S. (2015). School management, cultures of teaching and student outcomes: Comparing the cases of Finland and Sweden. *Teaching and Teacher Education*, 49, 168–177. <https://doi.org/10.1016/j.tate.2015.02.014>
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4), 347–363. <https://doi.org/10.1177/0734282911406661>
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 167–180. <https://doi.org/10.1080/10705511.2014.882658>
- Scheerens, J. (2016). *educational effectiveness and ineffectiveness: A critical review of the knowledge base* (1st ed.). Dordrecht, The Netherlands: Springer.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. <https://doi.org/10.3102/0034654307310317>
- Soh, K. (2014). Finland and Singapore in PISA 2009: Similarities and differences in achievements and school management. *Compare: A Journal of Comparative and International Education*, 44(3), 455–471. <https://doi.org/10.1080/03057925.2013.787286>
- van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00770>
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- 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(1), 4–70.
- Vieluf, S., Leon, J., & Carstens, R. (2010). Construction and validation of scales and indices. In *TALIS 2008 technical report* (pp. 131–206).
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, 63(3), 249–294. <https://doi.org/10.3102/00346543063003249>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

