# Teacher autonomy and responsibility variation and association with student performance in Didaktik and curriculum traditions

Armend Tahirsylaj

Published online: 16 Oct 2018.

Submit your article to this journal ☑

Article views: 724

View Crossmark data ☑

Routledge
Taylor & Francis Group

# Teacher autonomy and responsibility variation and association with student performance in Didaktik and curriculum traditions

Armend Tahirsylaj 🔘

Department of Education, Linnaeus University, Växjö, Sweden

**ABSTRACT**

The primary objective of the study was to empirically test theoretical claims made about differences between Didaktik and curriculum traditions concerning teacher autonomy (TA) and teacher responsibility (TR). It tests the hypothesis that TA and TR are higher among Didaktik than curriculum countries. The second objective was to explore associations of TA and responsibility measures with students' science performance? Nationally representative data from 2009 Programme for International Student Assessment (PISA), collected through a two-step random selection process were used. For TA individual items Mann–Whitney rank-sum test was employed, while a difference of proportion test was used for TR items to examine the differences. Hierarchical linear modelling (HLM) was used to examine association of TA and TR items with students' science performance in PISA 2009. Overall and contrary to the initial hypothesis, teachers in curriculum countries enjoy both more autonomy and responsibility than teachers in Didaktik countries, but differences were substantively weak. Furthermore, within-country associations of autonomy and responsibility measures with students' science performance were found in a few countries. Further research is recommended to address TA and responsibility and complexities that accompany them in current stakeholder-crowded school contexts.

## Introduction and purpose

To deepen understanding between German-based Didaktik and US-based curriculum traditions, an intensive dialogue took part during the 1990s. The spark for that intensified dialogue, after some earlier exchanges between two traditions starting from 1960s, came from the interest of curriculum scholars to better understand what the concept called 'Didaktik' was? (Hopmann, 2015; Hopmann & Riquarts, 2000). Theoretical exchanges that emerged from 'Didaktik meets curriculum' dialogue as well as extended theoretical contributions from many scholars to Didaktik and curriculum traditions dominating K-12 education system provide the framework for this study.

First, what is *Didaktik* theory? German *Didaktik* theory is central to curriculum, teaching and learning in Continental and Nordic Europe generally and German speaking world specifically, but is mostly unknown in the English-speaking world (Hopmann, 2007; Hudson, 2003; Kansanen, 1999; Westbury, Hopmann, & Riquarts, 2000). The German term *Didaktik* is well established in post-1990s English literature, and to avoid the use of more negatively loaded concept of didactics as teacher-centered and controlled instruction (Kansanen, 2002), the original spelling Didaktik is used throughout. In its original conceptualization, 'Didaktik is about how teaching can instigate learning,

CONTACT Armend Tahirsylaj ✉ armend.tahirsylaj@lnu.se 🖃 Department of Education, Faculty of Social Sciences, Linnaeus University, P G Vejdes väg, Växjö 351 95, Sweden

but learning as a content-based student activity not as swallowing a sermon or a monologue or otherwise one-sided distribution of knowledge by a teacher' (Hopmann, 2007, p. 113). The interpretation of Didaktik across Europe led to a variety of modes, but Hopmann points out that they shared three common aspects, '(a) the concept of Bildung, (b) the embedded differential of matter and meaning, and (c) a concept of the necessary autonomy of teaching' to address 'problems of order, sequence, and choice within their respective frames of reference' (Hopmann, 2007, p. 115). Bildung, in classical sense, encompasses the contents of assisting individuals to achieve their self-determination by developing and using their reason without others' guidance and acquiring the cultural objects of the world in which those individuals are born into and situated in (Hudson, 2003; Klafki, 2000).

## Statement of purpose

The primary purpose was to empirically test theoretical claims made about Didaktik and curriculum in previous literature. For example, one of the core claims in prior Didaktik/curriculum exchanges was that there is more teacher autonomy (TA) and professionalism in Didaktik systems than in curriculum systems (Hopmann, 2007; Westbury, 2000). This study explored this argument in depth to clarify whether the claim was supported by empirical data. Furthermore, the purpose was to understand and further extend cross-national and cross-cultural educational dialogues and exchanges between Didaktik and curriculum theorists, educational practitioners, researchers, and policy makers. A more intensive and extensive educational exchange between nations' scholars could potentially lead to first, a deeper and better understanding of respective pedagogical traditions, and second, better informed decision-makers and policy makers to help avoid challenges in policy development and program implementation.

The ongoing international large-scale assessments (ILSAs) provide rich student performance and student and school background data that make possible the dialogue and exchanges between and among countries all around the world pertaining to relevant and complex educational policy issues in K-12 education systems. While being considerate of their existing limitations, ILSAs offer opportunities for researchers to examine different questions of interest that could have not been possible to be addressed otherwise due to limitations in data collection capacities. To that end, another purpose of the study was to track whether and to what extent two sets of Didaktik (Austria, Germany, Denmark, Finland, Norway and Sweden) and curriculum (Australia, Canada, United Kingdom, Ireland, New Zealand and United States) countries align within own group, and how cohesive the Didaktik-curriculum division is, when such grouping is done on a list of multi-source criteria. In addition, the purpose was to investigate how key TA and TR measures based on data collected by PISA in 2009 are associated with student science achievement data for two sets of countries. Next, literature review, conceptual framework, research hypothesis, and a brief summary of education systems in 12 countries are elaborated on. Then, data, sample, variables and methods employed to conduct the study are presented. Next, main results and findings are noted, ending with discussion and conclusions.

## Literature review, conceptual framework, and context

### Cross-national student achievement and teacher characteristics studies

Research on cross-national comparisons as well as within individual countries with regard to predictors of student achievement is abundant. Three main groups of factors are usually associated with student achievement in past research, namely, family SES, teacher and school characteristics, and national level investments into education system. Hanushek and Woessmann (2010) found in a cross-national analysis of student achievement that institutional structures and the quality of teaching force accounted for a significant portion of international differences in the level and

equity of student achievement. Other researchers have found that late tracking and long-pre-school cycle contributed to equality of opportunity with regard to educational performance (Schütz, Ursprung, & Wößmann, 2008). On the other hand, teacher related characteristics, such as TA and responsibility are not often used as predictors of student achievement as they seem to be more discussed within the context of teacher education relying more on qualitative research methods (Smith, 2001).

In cases when teacher related variables were examined, especially in econometric methods-based research, almost always they are examined in the context of teacher quality and its relationship with student achievement (Hanushek, 1971; Hanushek & Rivkin, 2006; Rivkin, Hanushek, & Kain, 2005). A review of teacher quality related research revealed that the empirical research follows three distinct avenues, including, first, teacher salaries, since teachers constitute the largest budgetary line in schools; second, the extent to which specific teacher characteristics account for student achievement; and third, the total effects of teachers on student achievement without specifying any measurable teacher characteristic (Hanushek & Rivkin, 2006). This present study falls into the second line of teacher related research as it examined, at the analytical level, the relationship among teachers' characteristics – autonomy and responsibility variables – to student achievement.

Overall, comparative education scholarship is divided over usability and validity of cross-national education research due to national cultural and historical differences. Indeed, scholars participating in Didaktik-curriculum dialogue during 1990s noted that educational issues '[…] are always rooted in the particularities of national histories, of national habits, and of national aspirations' (Reid, 1998, pp. 11–12). In addition, considering the complexities of the twenty-first century advanced societies, even under heavy influence of internationalization and globalization processes, there always seem to be some incomparability within the comparable (or comparability within the incomparable) (Depaepe, 2002). Furthermore, a large body of literature is critical of international testing and OECD and PISA in particular for promoting a certain set of neo-liberal educational policies and approaches with sweeping negative implications for education systems globally (cf. Grek, 2009; Hopmann, Brinek, & Retz, 2007). However, constructive insights and deeper understandings may emerge while disentangling global and cross-national educational developments and practices and differences and similarities therein. It is argued that educational policy discourse has largely been standardized globally (Schriewer, 2000), still, cross-country educational differences in Didaktik and curriculum systems may serve as relevant potential research-based understandings of themselves and one another.

## Teacher autonomy and teacher responsibility

After a literature review, Smith (2001) identified three main definitions for TA: (1) (Capacity for) self-directed professional *action*, where [Teachers may be] 'autonomous in the sense of having a strong sense of personal responsibility for their teaching, exercising via continuous reflection and analysis … affective and cognitive control of the teaching process' (Little, 1995); (2) (Capacity for) self-directed professional *development*, where [The autonomous teacher is] 'one who is aware of why, when, where and how pedagogical skills can be *acquired* in the self-conscious awareness of teaching practice itself' (Tort-Moloney, 1997); and (3) *Freedom from control by others* over professional action or development. Smith (2001) argues that it is the third definition that dominates the literature on TA, and it is also adopted for the present study. Other scholars have discussed TA as a source of both power and control, by employing the concept of empowerment as a means to achieve TA, but which, at the same time, could potentially serve as a disciplinary tool of governance structures (Lawson, 2004). More recently, Frostenson (2012) suggests three dimensions of TA: (1) a *professional dimension* which regards autonomy as characteristic of teachers as a professional group, a *faculty or staff dimension* which emphasizes the autonomy of a school organization, including the principal and the whole teaching staff, and finally an *individual dimension*, which refers to autonomy that the individual teacher possesses. Furthermore, literature proposes an *institutional dimension to TA*, implying collective autonomy of the teaching profession, and a *service dimension to TA*, concerning

individual TA at classroom level and school level practices more broadly (Wermke & Höstfält, 2014). This study concerns the *Individual/Service* dimensions of TA defined here as teacher freedom from control by others over professional action or development, in particular with regard to the type of assessments used in school settings. Subsequently, the present study is not an attempt to develop a TA scale but rather draws from TA conceptual definitions to be able to test theoretical claims about differences in TA between Didaktik and curriculum education traditions. Studies that focus on developing TA scales target teacher audiences and employ item-based instruments and factor analysis to elicit dimensions related to TA. For example, Pearson and Hall (1993) identified two main traits of TA: *general autonomy* (relying on items concerning classroom standards of conduct and personal on-the-job discretion) and *curriculum autonomy* (based on items concerning selection of activities and materials and instructional planning and sequencing). The curriculum autonomy as defined by Pearson and Hall (1993) overlaps with the definition adopted in the present study regarding TA as freedom from control of others over professional action – professional action meaning decisions teachers make at the classroom context about teaching and assessment for example. Qualitative studies focusing on TA rely on teacher interviews and phenomenological and ecological theoretical framing to examine perceived teachers' autonomy in their working environment (cf. Erss, Kalmus, & Autio, 2016).

Regarding teacher responsibility (TR), the study relies on the definition provided by OECD (2009) in its assessment framework, where TR was taken to mean responsibility over decisions pertaining to school management, financial and instructional issues. In this study, items related to teacher's responsibility over assessment policies, textbooks used, course content and courses offered are utilized. Corcoran (1995) also refers to TR as their capacity to make curriculum- and assessment-related decisions. In another case, TR over curriculum and assessment matters is presented as part of the skill set of effective teachers (Stronge, 2007). Arguably, differences in definitions of TA and TR are not wide apart, and potentially at another macro-level of abstraction they could be theorized to imply the same concept, however, they are kept separate in the present study to relate more closely to the micro-level of specific indicators available in PISA dataset.

## Conceptual framework

As noted earlier, the study relies on pedagogical traditions of Didaktik and curriculum for theoretical framing. An extended comparative discussion of these two pedagogical traditions has been offered recently by Tahirsylaj, Niebert, and Duschl (2015), but in summary and in line with previous elaborations on Didaktik and curriculum (Hopmann, 2007; Westbury, 2000), Didaktik relies on the concept of *Bildung* and professional TA and responsibility and thus it is more teacher-oriented and content-focused. Curriculum on the other hand is more institution-oriented, methods-focused, and evaluation intensive, while also lacking the core concept of *Bildung*.

The research on how student, teacher and school factors are associated with student achievement is not new. The impetus was created with the publication of the report on *The Equality of Educational Opportunity* or as often referred to *Coleman report* (Coleman et al., 1966) that found that outside-of school factors had a larger effect on student achievement than school factors. Ever since, more and more studies have tried to either confirm or disconfirm those findings under different circumstances and at different contexts (Baker, Goesling, & LeTendre, 2002; Heyneman & Loxley, 1982). More than 400 similar studies have been undertaken since then (Hanushek & Rivkin, 2006).

However, most of these studies have explored the association of school and out-of-school factors to student achievement under sociological and economic frameworks. This study took a different approach as it analysed student, teacher and school characteristics under educational frameworks, i.e. under the Didaktik and curriculum pedagogical traditions. A rough sketch of this logic is given in Figure 1, where the diagram of logic for rationale is based on the assumptions that first, both student and teacher and school characteristics contribute to student achievement
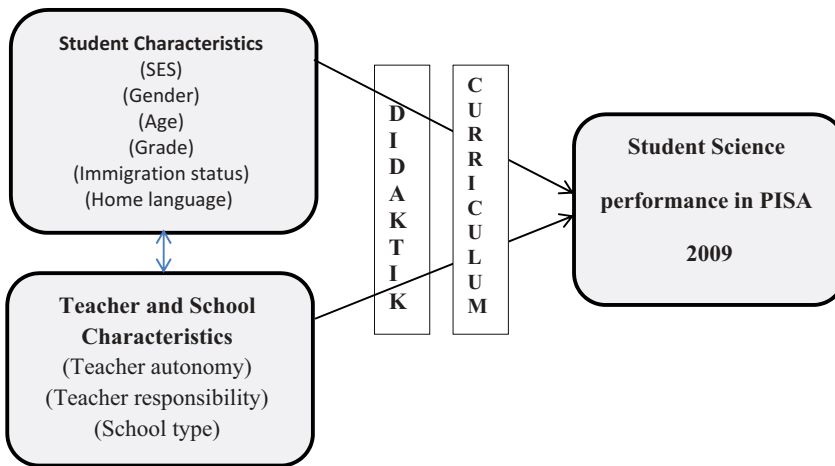
**Figure 1.** A diagram of logic rationale.

and are interrelated, and second, that those associations are mediated by the educational traditions in which they are embedded, in this case under Didaktik and curriculum. No prior research has used this framework to test student, teacher and school characteristics' associations with student achievement.

### Research questions and hypotheses

Didaktik and curriculum similarities and differences are multifaceted and complex, but the data available in PISA allowed for testing only a few of the claims that are made about Didaktik and curriculum, respectively. The study addressed two overarching questions that pertained to two types of analyses, i.e. descriptive and analytical. At the descriptive level, the main question was as follows:

(1) How do schools in Didaktik and curriculum countries compare across the core constructs representing teacher and school characteristics, namely teacher autonomy and teacher responsibility using PISA 2009 data?
   (a) Hypothesis: Schools in Didaktik countries will show higher levels of teacher autonomy, and teacher responsibility than curriculum countries.

At the analytical level, and given the findings from the first question, the study examined the relationship between teacher and school characteristics and student science test scores in PISA 2009. The question was as follows:

(2) What is the association of teacher and school characteristics with student science performance in PISA 2009 in representing Didaktik and curriculum countries?

Since the Didaktik-curriculum framework does not make any claim about the association of teacher and school characteristics with student achievement and since there was no other study that tested this relationship for Didaktik and curriculum countries using Didaktik and curriculum frameworks, there was no hypothesis developed for the second question, and it therefore, was exploratory in nature.

### A brief summary of education systems in sample countries

One overarching feature of the national education systems in the study pertains to a type of paradoxical stance – all countries are both seemingly similar and different at the same time! Examined from a macro-perspective they are all similar since none of them deviates from the 'norm' that has already been established for many decades, in the sense of organizing education along age-groups over a certain period of time. However, when each national system is explored in-depth, there are striking variations. For example, there are some clear differences between the Didaktik and curriculum countries with regard to tracking – tracking understood as separate organization of education into academic and vocational types of education. While K-12 education systems in curriculum countries are almost always comprehensive with little or no tracking at all, those in Didaktik countries vary from little tracking in Sweden and Denmark to 'heavy' tracking in Austria and Germany. Next, there are differences in terms of teacher education across Didaktik and curriculum traditions. For example, Didaktik countries recruit teachers from the highest ability student groups, train them longer before they are placed as teachers, and have extended periods of induction training. Curriculum countries are generally less selective in the recruitment process, but do require teachers to obtain masters degrees to enter the profession. Differences exist with regard to teacher professional development (TPD) as well. TPD is present in all countries, but while only recommended or not mandatory in Didaktik countries, TPD is required in curriculum countries, with exception of Ireland, where TPD is voluntary and up to teachers. Similarly, assessment systems constitute another point of divergence between Didaktik and curriculum systems, but there is also variation within systems as well. Overall, students in curriculum countries have to take far more external assessments during their K-12 education experience than in Didaktik countries. Lastly, concerning funding sources for public education systems there is less variation. Overall, national governments and municipalities share the costs to cover educational expenditures in almost all sample countries (Eurydice, 2015).

## Data and methods

### Data

This study utilized the data from the 2009 PISA survey. The Organisation for Economic Cooperation and Development (OECD) initiated its work on developing PISA surveys around mid-1990s and administered the first PISA survey in 2000 (OECD, 2002). The survey aimed to investigate whether 15-year olds across nations, irrespective of their grade, are ready for post-high-school life in terms of further work and education opportunities. PISA tests students' skills in three cognitive domains including mathematics, science and reading. It tests both knowledge and skills in any of these domains. PISA dataset fit best with the framework used for the study since the dialogue on Didaktik and curriculum was primarily focused on high-school level.

### Sampling and sample

A two-step sample selection process is followed in each PISA participating country. In the first step, the participating educational systems are broken down by key characteristics such as regions (federal states, provinces, cantons, etc.) and school types. Within these subdivisions, schools are then sampled at random. In the second step of the sampling process, 15-year-old students are sampled at random within the sampled schools (OECD, 2002). In all OECD countries, 15-year-old students are nearing the end of compulsory education.

In addition to testing students' knowledge and skills, PISA also collects background data by administering questionnaires to students, parents, and schools. Most often, school questionnaires are completed by school principals or administrators.

### Rationale for selection of countries for the study

A representative sample of 12 countries from Didaktik and curriculum traditions were examined in the study, a set of six for each. The so-called Germanic and Nordic countries were included in the Didaktik system and Anglo countries in the curriculum system. The rationale for the selection of countries was based on a number of criteria that justified this particular grouping.

First, there is the historical development of education systems. Didaktik emerged from the German-speaking world more broadly, and from nineteenth century Prussia (today's Germany) and it quickly spread to most of Continental Europe, and particularly to the Nordic countries, as one of the regions in proximity to German-speaking areas (Kansanen, 1999; Westbury et al., 2000). On the other hand, curriculum developed among Anglo-Saxon countries, primarily in United Kingdom and United States, and due to historical as well as linguistic ties permeates English-speaking countries around the world.

Second, there is a cultural element. Prior studies on world cultures, and more specifically Global Leadership and Organizational Behaviour Effectiveness Research Project (GLOBE) grouped world countries into 10 cultural clusters based on data from the surveys aimed at understanding organizational behaviour in respective societies (House, Hanges, Javidan, Dorfman, & Gupta, 2004). Didaktik countries below fall into two cultural clusters. Denmark, Finland, Norway and Sweden form the Nordic Europe Cluster while Austria and Germany fall into Germanic Europe cluster that also includes other regions and countries that speak German or Dutch in Europe. Curriculum countries selected for this study belong to the cluster called Anglo Cultures.

Third, there is empirical evidence from educational studies that the cultural clusters referred to above could be a potential way to differentiate these clusters in terms of within-country school differentiation (Zhang, Khan, & Tahirsylaj, 2015). Research findings from that study that used PISA data suggests that countries within Nordic, Germanic and Anglo cultures are similar to one another as to how much of the variance in student achievement can be explained by the differentiation among schools within respective countries.

Lastly and maybe most importantly, there is a practical element that pertains to the Didaktik-curriculum dialogue taking place during the 1990s. Two groups of scholars were involved. On one hand, there were scholars and researchers representing Didaktik that included both German and Nordic scholars, and on the other hand, there were curriculum experts that included scholars mainly from United Kingdom and United States (Gundem & Hopmann, 1998; Hopmann, 2015). Indeed, most of the meetings and symposia took place either in United States, Germany or Norway.

In terms of student achievement on PISA 2009 assessments, all 12 countries scored from high to higher as almost all of them scored above the OECD average of 500 points or were not statistically significantly below the OECD average, with exceptions of Austria, Sweden and Ireland. Also, economically, all the countries in both groups are OECD members, meaning they are part of developed countries, constituting some of the most developed economies in the world.

### Variables

All measures and variables for the study came from PISA 2009 dataset and Table 2 offers a detailed description of them. Table 1 provides school and student sample size for the countries in the study.

Prior to the data analysis, data were checked for any anomalies, in particular with regard to missing data. Overall, none of the key variables had more than 4% of data missing for $N = 9580$, while some of the school- and student-controls had a higher percentage of missing data. To address the first research question on variation of school-level items across curriculum and Didaktik countries, the complete case analysis was employed, i.e. only the observations with complete data were utilized for analysis. To address the second research question on relationships between TA and responsibility items with student science scores, two approaches were employed, namely mean substitution and dummy variable adjustment. The mean substitution method made possible using all information for all cases in the samples, while the missing dummy variable adjustment requires creation of a new

**Table 1.** School and student sample size for Didaktik and curriculum countries.

| | PISA 2009 | | |
|---|---|---|---|
| | School sample size | Overall student sample size | Average within- school sample size |
| **Didaktik countries** | **N = 1381** | **n = 32530** | |
| Denmark (DNK) | 285 | 5924 | 20.8 |
| Finland (FIN) | 203 | 5810 | 28.6 |
| Norway (NOR) | 197 | 4660 | 23.7 |
| Sweden (SWE) | 189 | 4567 | 24.2 |
| Austria (AUT) | 281 | 6590 | 23.4 |
| Germany (DEU) | 226 | 4979 | 22 |
| **Curriculum countries** | **N = 2285** | **n = 63450** | |
| Australia (AUS) | 353 | 14251 | 40.4 |
| Canada (CAN) | 978 | 23207 | 23.7 |
| Ireland (IRL) | 144 | 3937 | 27.3 |
| New Zealand (NZL) | 163 | 4643 | 28.5 |
| United Kingdom (GBR) | 482 | 12179 | 25.3 |
| United States (USA) | 165 | 5233 | 31.7 |
| **Total** | **3858** | **95980** | |

Source: OECD/PISA database.

variable indicating missing data for the given variable and including it in the statistical model when conducting analyses and the approach eliminates the variance in the dependent variable that is attributable to missing data (McKnight, McKnight, Sidani, & Figueredo, 2007). PISA dataset provides five plausible values for student achievement in each of the three content areas and final student weights and school weights to correct for the design effects were used when conducting statistical analysis. In this regard, the study does not use the scales and indices related to autonomy and responsibility measures already developed in the PISA dataset since the study focuses on results about teachers, while in the PISA dataset responses from the principal, the school's governing board, the regional or local education authorities or the national education authority are also used to construct the school responsibility indices (OECD, 2010).

## Descriptive statistics

Since the means of the main independent school-level variables (TA and TR items) constitute the focus of the first research question about variation, their means are presented in detail in Results section, while detailed descriptive statistics for other dependent and independent variables are available upon request.

## Methods

First, to address the first research question, Mann–Whitney rank-sum test was employed for TA individual items to examine the differences between the schools in two traditions as the items were measured on an ordinal scale, while a difference of proportion test was used for TR items to examine the differences. Second – given the results from the first question – to examine the relationship between teacher and school characteristics and student science test scores in PISA 2009 a hierarchical linear modelling (HLM) procedure was conducted. HLM was a more useful method for this study considering that PISA datasets have a nested structure of data with students nested in schools and schools nested in regions and/or countries, which arguably provides more precise estimates given that structure (Raudenbush & Bryk, 2002). Furthermore, HLM is preferred over simple ordinary least-squares (OLS) method since the latter assumes independence of observations, which is rather misleading in nested data where variance within group tends to remain dependent.

**Table 2.** Main study variables.

| Study measures | |
|---|---|
| Short description | Full description (original scale) |
| **Dependent variable** | |
| Science test scores | Five plausible values of science test scores in PISA 2009 |
| **Independent measures** | |
| **Level-1 (student) measures** | |
| Socio-economic status (SES) | Index of economic, social and cultural status created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI); the highest level of education of the student's parents, converted into years of schooling; the PISA index of family wealth; the PISA index of home educational resources; and the PISA index of possessions related to 'classical' culture in the family home |
| Gender | Are you female or male? (female = 1, male = 0) |
| Age | Age of the student |
| Grade | Grade compared to the modal grade in the country |
| First-generation immigrant | (first-generation immigrant = 1, 0 otherwise) |
| Second-generation immigrant | (second-generation immigrant = 1, 0 otherwise) |
| Home language other than the test language | (home language other than test language = 1, home language same as test language = 0) |
| **Level-2 (school) measures** | |
| Teacher autonomy | Generally, in your school, how often 15 year olds are assessed using:<br>a. standardized tests<br>b. teacher developed tests, and<br>c. teacher judgmental ratings.<br>(Scale: 1 = never, 2 = 1–2 times a year, 3 = 3–5 times a year, 4 = monthly, 5 = more than once a month) |
| Teacher responsibility | Regarding your school, who has considerable responsibility for the following tasks?<br>a. Establishing student assessment policies<br>b. Choosing which textbooks are used<br>c. Determining course content, and<br>d. Deciding which courses are offered<br>(1 = teachers, 0 = otherwise (principals, school governing board, regional or local education authority, national education authority) your school, who has considerable responsibility for the following tasks? |
| School type | Is your school a public or private school? (public = 1, private = 0) |

To develop the HLM models, first an unconditional model was run for each country using the dependent variable. Here is the specified equation for science achievement:

$$science_{ij} = \beta_{0j} + e_{ij} \tag{1}$$

Each school's intercept, $\beta_{0j}$, is then set equal to a grand mean, $\gamma_{00}$, and a random error $u_{0j}$:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2}$$

where $j$ represents schools and $i$ represents students with a given country.

Substituting (2) into (1) produces

$$science_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{3}$$

where

$\beta_{0j}$ is the mean science achievement for school j,

$\gamma_{00}$ is the grand mean for science achievement,

Var $(e_{ij}) = \theta$ is the within school variance in science achievement and

Var $(u_{0j}) = \tau_{00}$ is the between school variance in science achievement.

This model explains whether there is variation in student's standardized science scores across $j$ schools for the given country.

From here, a linear random-intercept model with covariates was set up. This model is an example of linear mixed effects model that splits the total residual or error into two error components. It starts with a multiple-regression model as follows:

$$Science\ scores_{ij} = \beta_1 + \beta_{2j}\ x_{2ij} + \ldots + \beta_p\ x_{pij} + \xi_{ij} \tag{4}$$

where $\beta_1$ is the constant for the model, while $\beta_{2j}\ x_{2ij}$ to $\beta_p\ x_{pij}$ represent covariates included in the given model. $\xi_{ij}$ is the total residual that is split into two error components:

$$\xi_{ij} \equiv u_j + e_{ij} \tag{5}$$

where $u_j$ is a school specific error component representing the combined effects of omitted school characteristics or unobserved heterogeneity. It is a random intercept or the level-2 residual that remains constant across students, while level-1 residual $e_{ij}$ is a student-specific error component, which varies across students $i$ as well as schools $j$. Substituting $\xi_{ij}$ into the multiple-linear regression model (4), we obtain the linear random-intercept model with covariates:

$$Science\ scores_{ij} = \beta_1 + \beta_{2j}\ x_{2ij} + \ldots + \beta_p\ x_{pij} + u_j + e_{ij} \tag{6}$$

where $\beta_{2j}\ x_{2ij}$ to $\beta_p\ x_{pij}$ represent the covariates included in the model, and they vary depending on how many covariates are included in a specific model. Three linear random-intercept models are fit for each country in the study, where the first model includes three level-2 covariates representing TA items: namely, the use of standardized tests, the use teacher-developed tests and the use of teacher judgments. The second model adds another four level-2 covariates representing TR items whether teachers were responsible for assessment policies, textbooks to be used, course content and courses offered. The final model is a full model that introduces one school-level covariates of school type (public vs. private) and a number of student level-1 covariates including SES, age, grade, immigration status (native vs. first generation vs. second generation), test language (native vs. another), and a dummy variable for gender, where female = 1 and male = 0, into the model with TA and responsibility items, and controlling for dummy missing variables. The same models are run for each of the 12 countries in the study.

## Results

### *Means of teacher autonomy measures across sample countries*

The results for means of TA measures, including use of standardized assessments, teacher-developed tests and teacher judgments across sample countries are presented in graph form in the following sections.

Figure 2 shows the mean use of standardized tests for each country, a variable coded on a scale from 0 to 4. On this scale, a higher mean indicates lower TA, as we expect less use of standardized tests in countries where teachers have more autonomy regarding employing student assessments instruments.

Based on Didaktik-curriculum framework, teachers in Didaktik countries were expected to enjoy higher autonomy, thus the mean use on standardized tests were expected to be lower. As the graph shows, there are more Didaktik countries on the upper side of the scale than curriculum countries. Within Didaktik countries themselves there is a clear separation between the four Nordic countries (Sweden, Denmark, Norway, and Finland) and those from Continental Europe (Austria and Germany) with Nordic countries having the highest means. The OECD average for this scale is at 1.19 and only two of the curriculum countries (New Zealand and United States) are above the OECD average and the other four below it. A Mann–Whitney two-sample rank-sum test was used to compare curriculum and Didaktik countries as two groups. The overall analysis resulted with Didaktik countries having a higher sum of ranks compared to the expected rank sums under the null hypothesis than curriculum countries and there was a significant difference of z = 3.306 with

## Use of standardized tests (by country)
### 0=Never, 1=1-2 a year, 2=3-5 a year, 3=monthly, 4=more than once a month
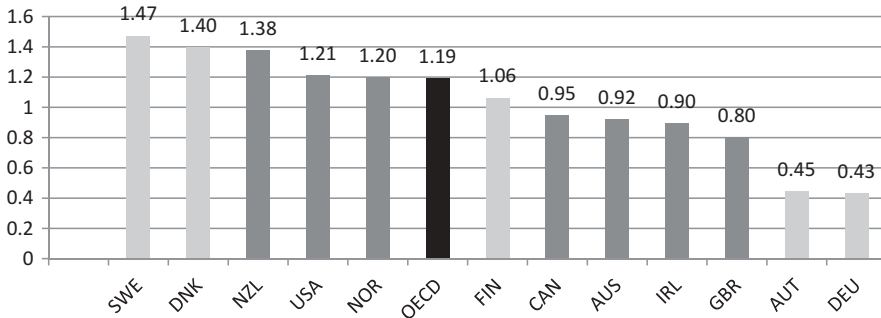#### Curriculum | Didaktik | OECD



**Figure 2.** Means of use of standardized tests by country.

$p < 0.001$. As a result, taken together Didaktik countries have a higher mean on the use of standardized tests than curriculum countries thus indicating that TA on average is higher in curriculum than Didaktik countries.

The next item part of TA construct was the use of teacher-developed tests. Here, a higher mean indicates higher TA as the locus of control for student assessment purposes lies with teachers. Figure 3 shows the results for curriculum and Didaktik countries. Here, the means for Didaktik countries are around the OECD average of 2.89, while those of curriculum countries are spread out on both ends of the continuum.

On average, apart from Ireland, United Kingdom and Denmark, all other countries have a mean close to 3, meaning teachers in those countries are reported to use tests developed by them on a monthly basis during the school year. In United States and Canada the mean is close to 4, indicating that in their schools teacher-developed tests are used more than once a month.

To test the difference between the two groups overall, the Mann–Whitney rank-sum test was employed. The results indicated that curriculum countries in total had a higher sum of ranks than

## Use of teacher-developed tests (by country)
### 0=Never, 1=1-2 a year, 2=3-5 a year, 3=monthly, 4=more than once a month
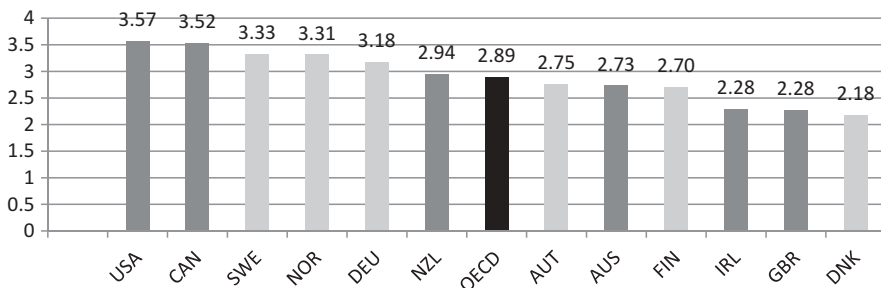#### Curriculum | Didaktik | OECD



**Figure 3.** Means of use of teacher-developed tests by country.

the expected sum of ranks, while for Didaktik countries the sum of ranks was lower than expected and there was a significant statistical difference of $z = -6.533$, with $p < 0.001$. Here, teachers in curriculum countries enjoy higher autonomy, however, if teacher-developed tests represent a form of preparation for students to take standardized tests than, indirectly, a higher mean on this item might indicate less TA.

The third item part of TA is the use of teacher judgments for ranking students as part of student assessment practices in respective countries. Figure 4 shows the means for each country, and here a clear separation between curriculum and Didaktik countries is observed with all Didaktik countries but Denmark positioned above the OECD average, and all curriculum countries but New Zealand falling below the OECD average. The Mann–Whitney rank-sum test confirmed the statistical difference of $z = 12.158$ at $p < 0.001$ between the curriculum and Didaktik countries, with Didaktik countries having higher sums of ranks than expected while curriculum countries' sums of ranks were lower than expected. In summary, examining the three previous graphs, it is observed that since all of the countries in the study are OECD member states, their average is around the OECD average.

### *Proportions of teacher responsibility measures across sample countries*

The results for proportions of TR measures across sample countries are presented in graph form in the following section.

Figure 5 shows the proportions of schools where teachers are reported to be responsible for making decisions about student assessment policies in respective countries. As per curriculum-Didaktik framework, the hypothesis was that more schools report teachers to be responsible in Didaktik than in curriculum countries. The graph shows that Austria (85.5%) and Ireland (84.0%) have the highest proportion of schools where teachers are responsible for assessment policies. United States has the lowest proportion of schools that have teachers who are responsible for assessment policies as reported by school principals with 44.3%. Most countries in the sample are above the OECD average. No clear separation is observed between curriculum and Didaktik countries. A difference-of-proportion test showed that Didaktik countries together had a mean of 0.69 (overall 69% of teachers in Didaktik countries are reported to be responsible for deciding about assessment policies) while curriculum countries had a mean of 0.74. The $z$-statistic was $z = -3.048$, $p < 0.01$ indicating that the difference between the two groups was statistically significant.
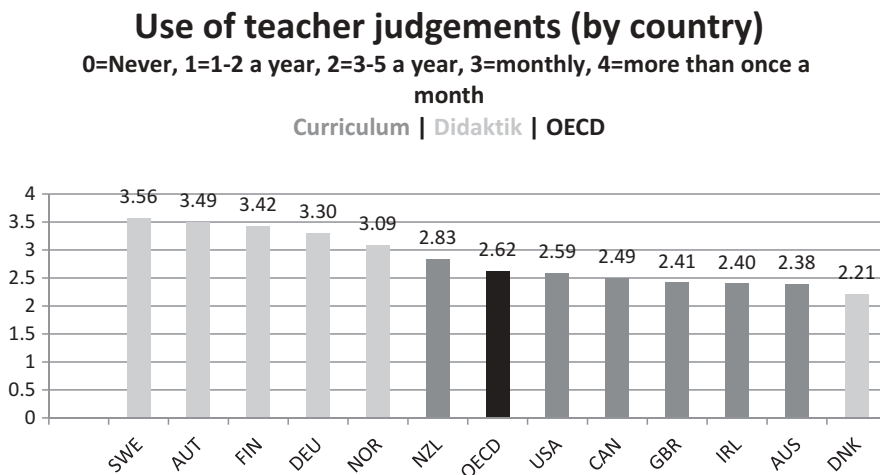


**Figure 4.** Means of use of teacher judgements by country.

## Proportion of teachers deciding about assessment policies (by country)
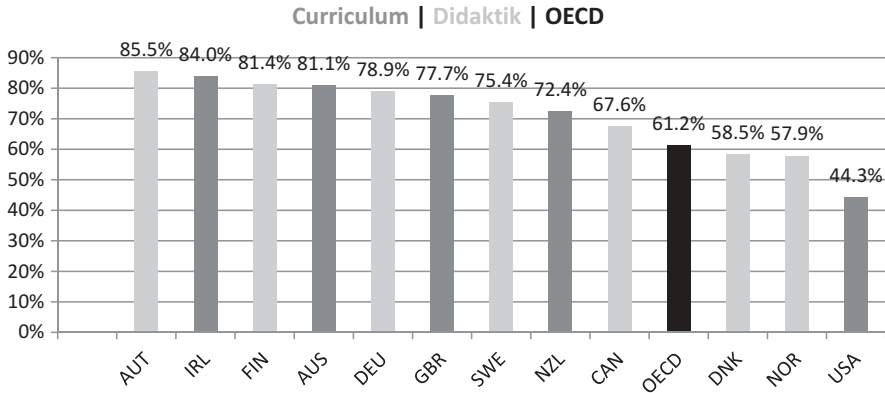
Curriculum | Didaktik | OECD



**Figure 5.** Proportion of teachers deciding about assessment policies by country.

Figure 6 shows the proportion of schools with teachers reported to be responsible for deciding about textbooks used in their respective countries. Overall, the proportions are quite high and above 60% for all countries, and all but United States are above OECD average of 67.3%. In United Kingdom, Sweden, Ireland, Finland, New Zealand and Australia close to a 100% of schools are reported to have teachers who decide about textbooks to be used. Overall, there is no clear separation between the two groups. This was also confirmed by the proportion-of-differences test, which showed that in both Didaktik and curriculum countries in their total school samples about 86% of teachers are reported to decide about textbooks used. The $z$-statistic $z = 0.245$, $p > 0.05$ indicated that the difference in proportions between the two groups was not statistically significant.

Figure 7 shows the proportion of schools reported by principals to have teachers who decide about course content in their respective countries. Again there is no clear separation of curriculum and Didaktik countries, and all countries in the sample are above the OECD average of 0.55. United Kingdom, New Zealand and Sweden have more than 90% of schools where teachers are reported

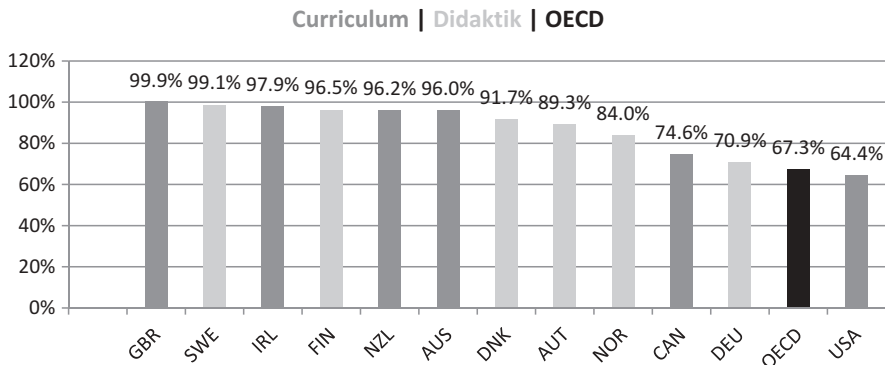## Proportion of teachers deciding about textbook use (by country)

Curriculum | Didaktik | OECD



**Figure 6.** Proportion of teachers deciding about textbooks to be used by country.

## Proportion of teachers deciding about course content (by country)
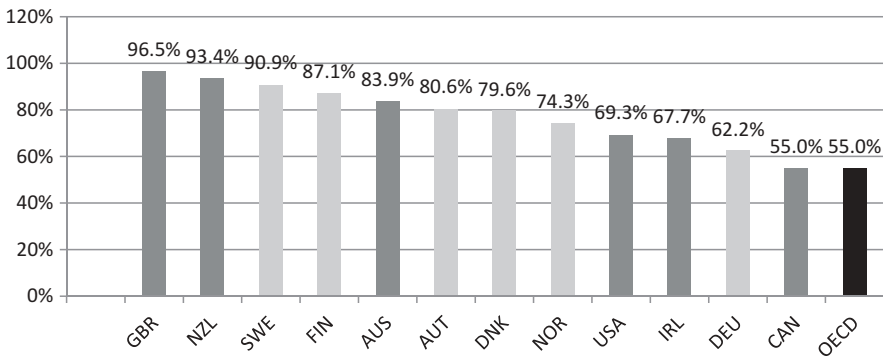
**Curriculum | Didaktik | OECD**



**Figure 7.** Proportion of teachers deciding about course content by country.

to decide about course content. In other countries, the range is from a low of 55% in Canada to a high of 87% in Finland. The proportion-of-differences tests showed that on average about 77% of teachers in Didaktik countries decide about course content compared to about 72% of teachers doing the same in curriculum countries. The $z$-statistic was $z = 3.619$ and $p < 0.001$ indicating that the difference in proportions between the two groups was statistically significant, with Didaktik countries having higher proportions.

Lastly, Figure 8 shows the proportion of schools where school principals reported that teachers decide about courses offered in their schools. The separation of Didaktik and curriculum countries is most visible with regard to the proportions between the two groups, where all Didaktik countries, apart from Finland, show that less than 50% of teachers decide about courses offered. Curriculum countries on the other hand show that the proportions vary from 55.6% in United States to 82.9% in United Kingdom. The OECD average for this item was 36.6% indicating that proportions in curriculum countries were significantly higher. Only Norway (21.5%) was below OECD average on this item. The

## Proportion of teachers deciding about courses offered (by country)
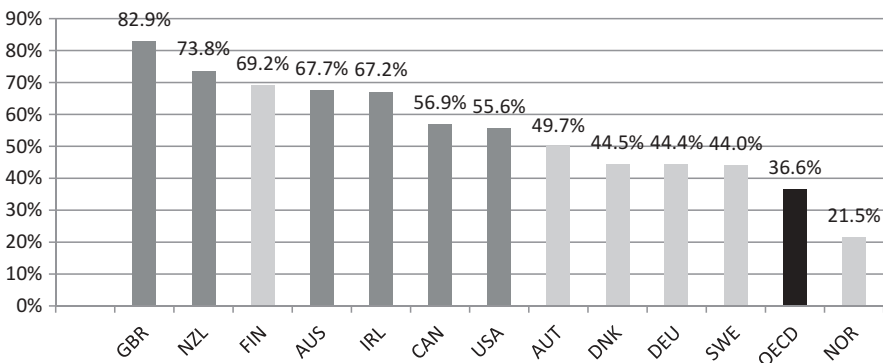
**Curriculum | Didaktik | OECD**



**Figure 8.** Proportion of teachers deciding about courses offered by country.

proportion-of-differences test showed the proportion of teachers in curriculum countries was higher at 0.66 while that of Didaktik countries was at 0.42, indicating that on average about 24% more teachers in curriculum countries are reported to decide about courses offered than their colleagues in Didaktik countries. The z-statistic was $z = -14.30$ and $p < 0.001$, indicating that the difference in proportions between the two groups was statistically significant.

Overall, out of four TR items, the proportion was significantly higher in curriculum countries on two of them (deciding about student assessment policies and courses offered), the proportion was statistically higher in Didaktik countries with regard to teachers who decide about course content, and there was no statistically significant difference between Didaktik and curriculum countries with regard to choosing textbooks to be used.

## Association of teacher autonomy and responsibility with student science performance in PISA 2009

The second research question of the study pertained to the 'so what' question, i.e. given the variation in TA and responsibility items across schools in curriculum and Didaktik countries, the objective was to examine whether those items are associated with student science performance in PISA 2009. Initially, a null model for each country was run to estimate the intra-class correlation (ICC), which represents the variation in science achievement across schools within every given country. The ICC results, which are consistent with those of previous studies (for example Zhang et al., 2015), varied from a low of 0.09 in Finland to highs of 0.57 in Austria and 0.65 in Germany. A low ICC of 0.09 in Finland to high ICCs of 0.57 in Austria and 0.65 in Germany suggests that schools in Finland performed about the same in PISA 2009 assessment, while in Austria and Germany there is a large variation in students' science performance across schools. These variations in Austria and Germany are usually attributed to the tracking orientations within K-12 educational systems. In the other 9 countries, the ICC range was from 0.12 in Norway to 0.29 in United Kingdom and United States.

Next, three linear random-intercept models were fit: first, one with TA items only; the second including both TA and TR items; and third, a full model that included both level-2 TA and TR items and level-2 and level-1 covariates. Detailed results for each curriculum and Didaktik country are available upon request, while Tables 3 and 4 below show summary results for the full model only (TA + TR + Level-2 and Level-1 covariates) that are presented in three sections – 'Fixed part', 'Random part' and 'Derived estimates'.

The main interest of the study pertains to the 'Fixed part' estimated regression coefficients as the objective is to examine the association of covariates to student science performance. Across curriculum countries, in the full model with level-1 and level-2 covariates, there are significant coefficients of TA and TR items in Ireland, New Zealand and United States. The coefficients of the use of standardized test ($\beta_2$) were mostly negative in the full model, but not significant in any of the curriculum countries. However, $\beta_2$ was negative and significant in the United States in first (−31.8) and second (−23.86) random intercept models that included only level-2 covariates but it lost significance in the full model after adding level-1 covariates. The coefficients for United States in the models with level-2 covariates indicate that the use of standardized tests is expected to decrease student science performance significantly in United States.

The coefficients for teacher-developed tests ($\beta_3$) were significant and positive only in New Zealand (8.25) and United States (29.6) in the full model controlling for other covariates and missing dummies among curriculum countries. The coefficient of teacher-developed tests for United States is particularly interesting as it is higher than SES coefficient (22.39) controlling for other covariates. At least in New Zealand and the United States, the monthly use of teacher-developed tests is estimated to increase student science performance. The United States had the highest mean in the variable of teacher-developed tests. The third item representing TA, the use of teacher judgments ($\beta_4$) was not significant in any of the curriculum countries.

**Table 3.** Maximum-likelihood estimates for science performance in curriculum countries.

| Fixed part | Australia | Canada | United Kingdom | Ireland | New Zealand | United States |
|---|---|---|---|---|---|---|
| | Est. | Est. | Est. | Est. | Est. | Est. |
| $\beta_1$ (_cons) | 725.86*** | 767.09*** | 353.52*** | 745.95*** | 424.62*** | 885.82*** |
| $\beta_2$ (stand. tests) | 1.06 | −1.36 | −2.97 | 1.67 | −2.92 | −12.98 |
| $\beta_3$ (teacher tests) | 3.62 | −0.86 | −2.01 | −7.28 | 8.25* | 29.60*** |
| $\beta_4$ (teach. judgments) | −0.57 | 0.17 | 5.03 | 2.66 | 0.99 | −2.12 |
| $\beta_5$ (assess. policies) | 1.83 | 0.15 | 10.15 | −5.17 | 1.36 | −10.99 |
| $\beta_6$ (text. used) | −0.73 | 2.78 | 23.11 | −48.21* | −106.37*** | 11.27 |
| $\beta_7$ (course content) | 4.43 | −2.59 | −4.47 | −11.06 | 95.72*** | −26.23 |
| $\beta_8$ (courses offered) | −2.43 | −3.60 | −11.18 | 1.87 | 15.49*** | 10.65 |
| $\beta_9$ (SES) | 32.85*** | 22.13*** | 30.83*** | 30.02*** | 38.52*** | 22.39*** |
| $\beta_{10}$ (gender) | −4.15*** | −6.69*** | −10.12*** | −2.42 | 7.17 | −14.70*** |
| $\beta_{11}$ (age) | −13.06*** | −13.14*** | 11.49*** | −10.76 | 6.19 | −26.63** |
| $\beta_{12}$ (grade) | 39.62*** | 48.53*** | 33.99*** | 19.80*** | 29.87*** | 43.06*** |
| $\beta_{13}$ (1st immigration gen.) | −4.59 | −11.22** | 5.75 | −3.28 | 0.25 | −4.33 |
| $\beta_{14}$ (2nd immigration gen.) | 2.85 | −13.47*** | 14.26* | 12.17 | −3.45 | −2.34 |
| $\beta_{15}$ (other test lang.) | −19.92*** | −12.10*** | −16.25* | −27.65** | −42.51*** | −6.72 |
| $\beta_{16}$ (public school) | −26.21*** | −25.94** | −46.59*** | −25.77** | −18.35** | −38.00** |
| **Random part** | mixed | mixed | mixed | mixed | mixed | mixed |
| Var. (random intercept) | 1188.83 | 1103.28 | 1489.69 | 1464.74 | 724.61 | 1297.04 |
| Var. (level-1 residual) | 6481.58 | 5124.95 | 5559.13 | 6033.58 | 6722.36 | 5008.22 |
| ICC(rho)/xtreg | 0.24 | 0.23 | 0.29 | 0.24 | 0.23 | 0.29 |
| Wald chi2(21^) | 1827.01 | 1498.44 | 806.47 | 1257.68 | 1678.6 | 1785.35 |
| Log likelihood | −83124.87 | −379026.79 | −623957.37 | −92189.75 | −57752.06 | −409343.4 |
| Students(schools) | 14251(353) | 23207(978) | 12179(482) | 3937(144) | 4643(163) | 5233(165) |
| **Derived estimates** | | | | | | |
| $R^2$ (total variance) | 0.23 | 0.19 | 0.21 | 0.15 | 0.31 | 0.28 |
| $R_1^2$ (level−1 var.) | 0.14 | 0.13 | 0.13 | 0.1 | 0.19 | 0.2 |
| $R_2^2$ (level−2 var.) | 0.51 | 0.39 | 0.41 | 0.32 | 0.71 | 0.49 |

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Controlling for dummy missing variables. '^' indicates degrees of freedom for Wald $\chi^2$d tests, which varied from country to country depending on how many variables were kept in the model.

Among TR items, textbooks used ($\beta_6$) was significant and negative in Ireland (−48.21), while textbooks used and course content ($\beta_7$) were significant and unexpectedly high in New Zealand, where if teachers were responsible for the textbooks to be used in schools then it was associated with a coefficient of −106.37 decrease of student performance. When teachers were responsible for the course content in New Zealand, it had a positive coefficient of 95.72, and courses offered variable ($\beta_8$) was also significant and positive (15.49) in New Zealand. When teachers had responsibility for assessment policies ($\beta_5$) it was not significant in any of the curriculum countries. Regarding level-1 covariates – SES ($\beta_9$), as expected, was significant and positive in all curriculum countries, while gender ($\beta_{10}$) was statistically significant and negative in five countries – Australia, Canada, United Kingdom, Ireland, and United States – significant and positive only in New Zealand, in the model controlling for other covariates. From among the remaining level-1 covariates, age ($\beta_{11}$) was significant and negative in Australia, Canada, and United States, significant and positive in United Kingdom, and not significant in Ireland and New Zealand. Grade ($\beta_{12}$) was significant and positive in all six curriculum countries, while variables representing first immigration generation students ($\beta_{13}$) and second immigration students ($\beta_{14}$) were significant and negative in Canada, significant and positive in United Kingdom, and not significant in other countries. The results suggest that students with immigration background are expected to perform better than native students in United Kingdom and worse than native students in Canada. When the home language of the student is different than test language ($\beta_{15}$) was significant and negative in all countries but United States. Lastly, level-2 covariate representing public schools ($\beta_{16}$) was significant and negative in all six curriculum countries, indicating that within each respective country, students attending public schools are expected to perform worse than those in private schools, as per the model employed and covariates used.

The 'Random part' section of the table shows the unexplained variance for level-2 (random intercept) and unexplained variance for level-1 after controlling for explanatory variables and missing dummies. The ICC shows variation across schools in science performance, while Wald $\chi^2$ and log-likelihood estimates are tests of the null hypothesis that coefficients used in the model are equal to zero. The number of students and schools per country is also given under the 'Random part'. Lastly, coefficients of determination or '$R^2$' are provided in the 'Derived estimates' section. Total $R^2$ indicates how much variation the model of interest explains. First, total variance of the null model and of two out of three target models (TA, and full model) is calculated as follows:

$$R^{2\,(total)} = \frac{total\ null\ model\ variance - total\ model\ of\ interest\ variance}{total\ null\ model\ variance} \tag{7}$$

$R^2$ were not calculated for the TA-TR model as it didn't differ substantively from TA model. Then, following Raudenbush and Bryk (2002) suggestion, the same procedure to partition the between-school and within-school level variance is applied for calculating $R^2$ for level-1 and level-2.

Table 4 provides maximum-likelihood estimates for science performance in Didaktik countries controlling for the covariates and missing dummies based on the same models as for curriculum countries. Among, the variables of interest representing TA, the use of standardized tests ($\beta_2$) was negative in all Didaktik countries but Denmark and Finland. In Finland, the coefficient for use of standardized tests was significant and positive (16.74). However, it was significant and negative in Austria (−19.38) and Germany (−30.83), indicating that the use of standardized tests in these two countries was associated with lower student science performance in PISA 2009 after controlling for covariates. The use of teacher-developed tests ($\beta_3$) was significant only in Norway (6.07), while the use of teacher judgments ($\beta_4$) was significant only in Sweden (7.07).

Regarding TR items, when teachers were responsible for assessment policies ($\beta_5$) it was negative and statistically significant in Austria (−30.97), positive and significant in Finland (12.90) and Norway (11.63), and it was not significant in other countries. When teachers were responsible for textbooks to be used ($\beta_6$) and course content ($\beta_7$) they were not significant in any of the Didaktik countries after controlling for covariates, while when teachers were responsible for courses offered

**Table 4.** Maximum-likelihood estimates for science performance in Didaktik countries.

| Fixed part | Austria | Germany | Denmark | Finland | Norway | Sweden |
|---|---|---|---|---|---|---|
| | Est. | Est. | Est. | Est. | Est. | Est. |
| $\beta_1$ (_cons) | 693.63*** | 763.21*** | 650.94*** | 685.43*** | 199.60*** | 572.18*** |
| $\beta_2$ (stand. tests) | −19.38** | −25.97* | 4.42 | 16.74*** | −2.89 | −10.41 |
| $\beta_3$ (teacher tests) | 1.17 | 0.62 | 1.64 | −4.82 | 6.07* | 2.5 |
| $\beta_4$ (teach. judgments) | 8.03 | −1.48 | 1.99 | 3.41 | 0.12 | −7.07* |
| $\beta_5$ (assess. policies) | −30.97* | −5.05 | 7 | 12.90* | 11.63* | 8.35 |
| $\beta_6$ (text. used) | 0.12 | 11.57 | 14.74 | −0.03 | −2.32 | −15.72 |
| $\beta_7$ (course content) | −3.55 | −6.97 | 1.14 | −0.99 | −0.58 | 0.34 |
| $\beta_8$ (courses offered) | 5.09 | −36.16** | −6.03 | −0.67 | 14.28 | 2.65 |
| $\beta_9$ (SES) | 11.85*** | 10.26*** | 28.47*** | 27.03*** | 30.24*** | 32.84*** |
| $\beta_{10}$ (gender) | −25.71*** | −20.37*** | −15.15*** | 13.98*** | 4.48 | 2.7 |
| $\beta_{11}$ (age) | −10.45* | −13.35*** | −10.94 | −9.68 | 15.90** | −1.24 |
| $\beta_{12}$ (grade) | 39.50*** | 37.32*** | 40.19*** | 39.17*** | 14.15 | 57.10*** |
| $\beta_{13}$ (1st immigration gen.) | −21.21*** | −12.78** | −37.28*** | −29.22 | −7.68 | −30.26** |
| $\beta_{14}$ (2nd immigration gen.) | −30.58*** | −18.48*** | −34.82*** | −27.88* | −19.63 | −28.67*** |
| $\beta_{15}$ (other test lang.) | −12.91** | −16.12*** | −22.99*** | −35.76*** | −47.46*** | −15.07 |
| $\beta_{16}$ (public school) | −13.04 | −23.27 | −2.17 | −17.16 | 14.77 | −21.31 |
| **Random part** | mixed | mixed | mixed | mixed | mixed | mixed |
| Var. (random intercept) | 3073.84 | 4332.85 | 618.4 | 410.45 | 624.45 | 873.57 |
| Var. (level-1 residual) | 3614.06 | 2634.14 | 5329.18 | 5707.73 | 5653.72 | 6361.53 |
| ICC(rho)/xtreg | 0.57 | 0.65 | 0.21 | 0.09 | 0.12 | 0.17 |
| Wald $\chi^2$(22^) | 2057.46 | 1222.86 | 1312.38 | – | – | 909.21 |
| Log likelihood | −259127.77 | −1454987.5 | −185910.1 | −121355.85 | −123893.3 | −23090.04 |
| Students(schools) | 6590(282) | 4979(226) | 5924(285) | 5810(203) | 4660(197) | 4576(189) |
| **Derived estimates** | | | | | | |
| $R^2$ (Total variance) | 0.33 | 0.31 | 0.31 | 0.15 | 0.14 | 0.21 |
| $R_1^2$ (level-1 var.) | 0.16 | 0.25 | 0.22 | 0.12 | 0.12 | 0.17 |
| $R_2^2$ (level-2 var.) | 0.46 | 0.35 | 0.66 | 0.39 | 0.3 | 0.44 |

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. '–' indicates estimate not reported by model. Controlling for dummy missing variables. '^' degrees of freedom for Wald $\chi^2$'d tests varied from country to country depending on how many variables were kept in the model.

($\beta_8$) it was significant only in Germany with a negative coefficient of −36.16. The same as in curriculum countries, SES ($\beta_9$) was significant and positive in all Didaktik countries, while gender ($\beta_{10}$) was significant and negative in Austria, Germany and Denmark, significant and positive in Finland, and not significant in Norway and Sweden.

From the rest of level-1 covariates, age ($\beta_{11}$) was significant and negative in Austria and Germany, significant and positive in Norway, and not significant in Denmark, Finland and Sweden. Grade ($\beta_{12}$) was significant and positive in all countries but Norway. The variable of the first immigration generation students ($\beta_{13}$) was significant and negative in all Didaktik countries but Finland and Norway, where it was not significant. The variable of second immigration students ($\beta_{14}$) was significant and negative in all countries but Norway. Differently from curriculum countries, the results suggest that students with immigration background are expected to perform worse than native students in the majority of Didaktik countries, and in no country they are expected to performer better than native students.

When the home language of the student is different than test language ($\beta_{15}$) was significant and negative in all countries but Sweden. Lastly and interestingly, the level-2 covariate representing public schools ($\beta_{16}$) was not significant in any of the Didaktik countries, indicating that there are no significant differences in student science performance between public and private schools within each respective Didaktik country, as per the model and covariates used in the study.

The 'Random part' provides the unexplained variance for the random intercept and level-1 residual, ICC values, Wald $\chi^2$d and log-likelihood estimates, and the number of students and schools. The $R^2$ values, under Derived estimates, for Didaktik countries are lower than for curriculum countries with exception of Denmark, suggesting that the random-intercept models were a better fit for curriculum than for Didaktik countries.

## Discussion and conclusions

The findings for the first research question were mixed, since taken together as one sample, teachers in schools in Didaktik countries do not necessarily enjoy more autonomy than those in curriculum countries, while teachers in schools in curriculum countries seem to enjoy more responsibility than those in Didaktik countries. Also, the variation was not as large as expected. The differences between Didaktik and curriculum countries were statistically significant, but substantively weak. However, TR was high across all countries implying that overall teachers in schools in the countries in the sample enjoy responsibility when it comes to decision-making concerning their professional classroom practices. With regard to TA, the frequent use of external standardized tests in most sample countries suggests that teachers are not free of control from others over their professional action concerning assessment instruments used in their classroom. This was a contradictory finding since on one hand, school leaders reported that teachers overwhelmingly make decisions about assessment policies, while at the same time there is a high frequency of standardized tests. Relying on Lawson's (2004) theorization then, the use of standardized tests serves more as an external control mechanism that diminishes TA.

Furthermore, one critical question remains as to why there were no substantive differences between TA and responsibility practices across schools in Didaktik and curriculum countries in PISA 2009, when the theory strongly suggested that would be the case. One plausible explanation for the lack of differences could be the global trends in education promoted by highly influential global education players such as OECD with PISA study or International Association for the Evaluation of Education (IEA) with TIMSS. The sample countries in the study have been consistently participating in international assessments during the past 30 years, and, if we take Germany as an example, they have introduced educational policies to align their educational systems with international assessments such as PISA. Same as when what gets tested what gets taught at the classroom level, what gets tested in international assessments drives what countries do in order to keep up with the ranking race associated with international assessments. As a result, while the rhetoric and

theory behind curriculum and Didaktik frameworks may indeed be valid, the school practices follow a pattern of their own that is more international than national – where the borders between local and global become thin and school practices around the world become similar. Ultimately, there is more variation in school practices within countries than between countries. Therefore, while the study relied on clearly theoretical divide between curriculum and Didaktik, it revealed practice convergence in TA and responsibility. To some extent, the study adds evidence to the isomorphic nature of educational practices around the world and further expansion of world culture of schooling, where countries continue to adopt similar practices in hopes to get similar better results (Baker, 2014; Baker & LeTendre, 2005). Subsequently, the results of the study challenge the Didaktik/curriculum dichotomy as per PISA data analysed here.

Findings from the analytical models of the second research question provide evidence for the predictive power, or lack thereof, of TA and responsibility items in explaining within-country students' science performance. Relationships between autonomy and responsibility items were only found in a small number of countries using within-country models. Indeed, in the majority of countries representing both groups of countries, the analytical models did not identify any association of TA and TR measures with science performance. One plausible explanation for this finding, at least with regard to TR measures, could be related to the fact TR items are only a small part of total proportion of responsibility shared among numerous stakeholders affecting school policies and practices as part of school autonomy. Teachers represent one group of stakeholders who are involved in making decisions about such important school issues as assessment policies, textbooks to be uses, course content and courses offered. The same justification might apply to TA. OECD itself emphasizes school autonomy as one of the features of educational systems that produce better performance results, but only when there is collaboration between school principals and teachers within the school (OECD, 2013). Therefore, from this perspective, TA is also part of the broader definition of school autonomy that includes both other players (e.g. school principals) and other practices (e.g. autonomy in formulating school budget). To that end, the development of new and better measures and methods that capture TA and TR and that are distinguished from school autonomy and responsibility measures would be valuable in examining the relationship of these constructs with students' performance.

In conclusion, cross-national educational research and subsequent exchanges in education policy have been around since 1950s, when some of these 12 countries were major players in such initial undertakings. Indeed, the first international comparative education organization emerged after a visit of Torsten Husén from the University of Stockholm at the University of Chicago in the mid-1950s, where among others he met with the famous Benjamin Bloom '[…] whose view was that the whole world should be seen as a single educational laboratory' (Baker, Lee, & Heyneman, 2003, p. 332). We are witnesses of how far the world has come to become the educational laboratory that Bloom envisioned! However, the exchange of cross-continental educational ideas, at least between United States and Germany, had occurred intensively more than a century ago. Reminiscent of that period, one of the present curriculum theorists writes, 'One hundred years ago, Americans travelled to Germany […] to study concepts of education. It seems to me it is time again to selectively incorporate German concepts in North American practices of education' (Pinar, 2011, p. xiv). Pinar primarily refers to elements of Bildung as German concepts that should be incorporated into American education. This is not to suggest that Germany should be the go-to destination to import education policies or concepts from, nor that the Didaktik model should be fully adopted, as the findings of the study did not suggest an advantage of one tradition over the other. However, the countries in this study, as well as any other country that participates in international assessments in particular, have the advantage of benchmarking themselves against the best performing nations in the world, and can eclectically pick and choose – with caution and consideration for local contexts – educational policies and practices to incorporate in their respective educational systems. Ideally, those policies should

potentially contribute to empower teachers to make autonomous professional decisions in their work, which by implication, results on self-directed, empowered learners – much needed for the present and imminent challenges of our and their times.

### Limitations of the study and further research

While the findings of this study may be generalized to specific countries that were examined more thoroughly here as well as to schools within them, a number of limitations need to be recognized. First, the data used here is cross-sectional and no causality is meant with the findings. Also, the study relied on secondary-data analysis; therefore it was not possible to control what data was collected. The variables used were derived from those available on the PISA 2009 dataset, which served as proxies to TA and TR constructs that the study examined. Variables that were selected attempted to narrow down teacher-specific items only. The analytical models employed were specifically focused on relationships of TA and responsibility measures with student science performance, controlling for two student-specific variables, and the list of variables used for predicting science achievement is selective and in no way exhaustive. A future research project that could potentially tackle some of study's limitations could focus on TR. The study found that, overall, there is a high proportion of teachers who are responsible for crucial teaching and learning matters. On the other hand, only a few of responsibility measures were found to be significantly associated to students' science performance. One way to explore this issue could be to view responsibility from a broader perspective, by examining the proportion of responsibility of other educational stakeholders, such as school principles and other local or national authorities, who are responsible for making decisions relevant to important issues such as assessment policies, text-books to be used, course content and courses offered. Furthermore, exploring the interactions of these stakeholders and how responsibility for decision-making is shared could provide further valuable insights as to how teachers are positioned in prevailing complex hierarchical decision-making structures within national educational systems. Finally, applying some of the more recent quasi-experimental quantitative research methods such as propensity score matching (PSM) or regression discontinuity (RD) and using other datasets in addition to PISA, such as Trends in International Mathematics and Sciences Study (TIMSS) or Programme for International Reading Literacy Study (PIRLS), wherever applicable, could provide more precise estimates of teacher-specific effects on student performance in ILSAs.

### Disclosure statement

No potential conflict of interest was reported by the author.

### Notes on contributor

*Armend Tahirsylaj* is a post-doctoral fellow of education in the Department of Education and Teachers' Practice, Linnaeus University, Växjö, Sweden. His research interests primarily include curriculum studies, Didaktik, education policy and theory, teacher education, international large-scale assessments, and international comparative education.

### ORCID

Armend Tahirsylaj 🆔 http://orcid.org/0000-0002-3902-0338

### References

Baker, D. P. (2014). *The schooled society: The educational transformation of culture*. Stanford, CA: Stanford University Press.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley Effect" on mathematics and science achievement. *Comparative Education Review*, *46*(3), 291–312.

Baker, D. P., Lee, J., & Heyneman, S. P. (2003). Should America be more like them? Cross-national high school achievement and US policy. *Brookings Papers on Education Policy*, 309–338.

Baker, D. P., & LeTendre, G. K. (2005). *National differences, global similarities: World culture and the future of schooling*. Stanford, CA: Stanford University Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Corcoran, T. B. (1995). *Transforming professional development for teachers. A guide for state policymakers*. Washington, DC: National Governors Association.

Depaepe, M. (2002). A comparative history of educational sciences: The comparability of the incomparable? *European Educational Research Journal*, *1*(1), 118–122.

Erss, M., Kalmus, V., & Autio, T. H. (2016). "Walking a fine line." Teachers' perception of curricular autonomy in Estonia, Finland, and Germany. *Journal of Curriculum Studies*, *48*(5), 589–609.

Eurydice. (2015). Welcome to Eurydice. Retrieved from https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Main_Page.

Frostenson, M. (2012). Lärarnas avprofessionalisering och autonomins mångtydighet [The deprofessionalisation of teachers and the ambiguity of autonomy]. *Nordiske Organisasjonsstudier*, *2*, 49–78.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23–37.

Gundem, B. B., & Hopmann, S. (1998). *Didaktik and/or curriculum*. New York: Peter Lang.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, *61*(2), 280–288.

Hanushek, E. A., & Rivkin, S. G. (2006). Chapter 18 Teacher quality. E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 2, pp. 1051–1078). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574069206020186

Hanushek, E. A., & Woessmann, L. (2010). *The economics of international differences in educational achievement*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w15949

Heyneman, S. P, & Loxley, W. A. (1982). Influences on academic achievement across high and low income countries: A re-analysis of iea data. *Sociology of Education*, *55*(1), 13-21. doi:10.2307/2112607

Hopmann, S., & Riquarts, K. (2000). Starting a dialogue: A beginning conversation between the Didaktik and curriculum traditions. In I. Westbury, S. Hopmann, & K. Riquarts (Eds.), *Teaching as a reflective practice: The German Didaktik tradition* (pp. 3–11). Mahwah, NJ: Lawrence Erlbaum.

Hopmann, S. (2007). Restrained teaching: The common core of Didaktik. *European Educational Research Journal*, *6*(2), 109–124.

Hopmann, S. (2015). 'Didaktik meets curriculum' revisited: Historical encounters, systematic experience, empirical limits. *Nordic Journal of Studies in Educational Policy*, *2015*(1), 27007.

Hopmann, S. T., Brinek, G., & Retzl, M. (Eds.). (2007). *PISA zufolge PISA – PISA according to PISA: Hält Pisa, was es verspricht? – Does PISA keep what it promises*. Vienna and Berlin: Lit Verlag.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 cultures*. Thousand Oaks, CA: Sage.

Hudson, B. (2003). Approaching educational research from the tradition of critical- constructive Didaktik. *Pedagogy, Culture and Society*, *11*(2), 173–187.

Kansanen, P. (1999). The Deutsche Didaktik and the American research on teaching. In B. Hudson, F. Buchberger, P. Kansanen, & H. Seel (Eds.), *Didaktik/Fachdidaktik as science(-s) of the teaching profession?* (pp. 21–36). Umeå, Sweden: Thematic Network of Teacher Education in Europe.

Kansanen, P. (2002). Didactics and its relation to educational psychology: Problems in translating a key concept across research communities. *International Review of Education*, *48*(6), 427–441.

Klafki, W. (2000). The significance of classical theories of Bildung for a contemporary concept of Allgemeinbildung. In I. Westbury, S. Hopmann, & K. Riquarts (Eds.), *Teaching as a reflective practice: The German Didaktik tradition* (pp. 85–107). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lawson, T. (2004). Teacher autonomy: Power or control? *Education 3-13*, *32*(3), 3–18.

Little, D. (1995). Learning as dialogue: The dependence of learner autonomy on teacher autonomy. *System*, *23*(2), 175–181.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York, NY: Guilford Press.

OECD. (2002). *Programme for International Student Assessment – PISA 2000 technical report*. Paris: Author.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris: Author.

OECD. (2010). *PISA 2009 results: What makes schools successful? – Volume IV*. Paris: Author.

OECD. (2013). *Lessons from PISA 2012 for the United States, strong performers and successful reformers in education*. Author. doi: 10.1787/9789264207585-en

Pearson, L. C., & Hall, B. W. (1993). Initial construct validation of the teaching autonomy scale. *Journal of Educational Research*, *86*(3), 172–178.

Pinar, W. F. (2011). *The character of curriculum studies: Bildung, currere, and the recurring question of the subject*. New York, NY: Palgrave Macmillan.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Reid, W. A. (1998). Systems and structures of myths and fables? A cross-cultural perspective on curriculum content. In B. B. Gundem & S. Hopmann (Eds.), *Didaktik and/or curriculum: An international dialogue* (pp. 11–27). New York: Peter Lang.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Schriewer, J. (2000). *System and interrelationship networks: The internationalization of education and the role of comparative inquiry*. In T. Popkewitz (Ed). *Educational Knowledge* (pp. 305–343). Albany, NY: SUNY Press.

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of opportunity. *Kyklos*, *61*(2), 279–308.

Smith, R. C. (2001). Teacher education for teacher-learner autonomy. *Language in Language Teacher Education*. Retrieved from http://homepages.warwick.ac.uk/~elsdr/Teacher_autonomy.pdf

Stronge, J. H. (2007). *Qualities of effective teachers*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Tahirsylaj, A., Niebert, K., & Duschl, R. (2015). Curriculum and Didaktik in 21st century: Still divergent or converging? *European Journal of Curriculum Studies.*, *2*(2), 262–281.

Tort-Moloney, D. (1997). *Teacher autonomy: A Vygotskian theoretical framework*. CLCS Occasional Paper No. 48. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED412741

Wermke, W., & Höstfält, G. (2014). Contextualizing teacher autonomy in time and space: A model for comparing various forms of governing the teaching profession. *Journal of Curriculum Studies*, *46*(1), 58–80.

Westbury, I. (2000). Teaching as a reflective practice: What might Didaktik teach curriculum? In I. Westbury, S. Hopmann, & K. Riquarts (Eds.), *Teaching as a reflective practice: The German Didaktik tradition* (pp. 15–39). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Westbury, I., Hopmann, S., & Riquarts, K. (Eds.). (2000). *Teaching as a reflective practice: The German Didaktik tradition*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Zhang, L., Khan, G., & Tahirsylaj, A. (2015). Student performance, school differentiation, and world cultures: Evidence from PISA 2009. *International Journal of Educational Development*, *42*, 43–53.