Nordic Journal of Comparative and International Education (NJCIE)

Trends in Teacher Monitoring Methods across Curriculum and Didaktik Traditions: Evidence from three PISA wave

Armend Tahirsylaj[1]

Associate Professor of Education, Department of Teacher Education, Faculty of Social and Educational Sciences, Norwegian University of Science and Technology (NTNU)

Abstract

The main objectives of this study were to examine trends in teacher monitoring methods (TMMs) among a representative set of 12 curriculum and didaktik countries, using data from PISA 2009, 2012, and 2015, and the association of TMMs with students' reading, mathematics, and science performance accordingly. Curriculum and didaktik education traditions frame the study theoretically, while quantitative research methods are used, consisting of a two-sample difference of proportion test and hierarchical linear modelling. The findings suggest that across the PISA waves, the control over teachers is growing across all countries and in all three subject domains and four TMMs. However, the proportion of students in schools where any of the TMMs are used is higher and more statistically significant for curriculum than for didaktik countries. Student tests, teacher peer review, and principal observation are much more common TMMs than external inspector observation across all countries. Nevertheless, the use of external inspector observation is very low in several didaktik countries, and in the case of Finland almost inexistent. Results for Sweden seem to be over-reported as in previous survey work it was found that teacher self-assessment is the most common TMM, however, teacher self-assessment is not a variable included in the PISA survey. The results from within-country hierarchical linear models (HLMs) of associations of TMMs with students' reading (2009), mathematics (2012), and science (2015) performance in PISA show mixed, and at times relatively large, effects from country to country and across three PISA waves, and interestingly the associations had diminished by PISA 2015. Adding a more diverse set of questions to PISA contextual questionnaires is warranted for results to be more meaningful and representative across more countries.

Keywords: teacher monitoring methods; accountability; curriculum; didaktik; international comparative education; PISA

---

[1] Corresponding author: armend.tahirsylaj@ntnu.no

Introduction and purpose

While teacher evaluation methods have attracted researchers' interest, primarily in the context the United States, as part of value-added models for accountability measures (e.g. Carnoy & Loeb, 2002; Darling-Hammond et al., 2012; Harris et al., 2014), no prior research has examined the trends in teacher monitoring and evaluation methods and their association with student performance from a comparative international perspective. With more intensive control mechanisms in place across countries, teachers are being exposed to pressure both internally through school mechanisms and externally through state authorities to 'produce' high results in the form of student test scores in locally, nationally, or internationally developed and administered tests. As a result, it is important to explore whether the enacted teacher monitoring methods are actually 'producing' the expected higher results. In this study, data from the Programme for International Student Assessment (PISA) 2009, 2012, and 2015 are examined from the curriculum and didaktik[2] perspectives as well as from that of PISA as a normative tool.

The study addressed two main research questions: (1) How do curriculum and didaktik traditions compare across four teacher monitoring methods (TMMs) using PISA 2009, 2012, and 2015 data?; and (2) What is the association of the four TMMs with students' reading, mathematics, and science performance in PISA 2009, 2012, and 2015, respectively, across a representative sample of 12 didaktik and curriculum countries? Teacher monitoring methods are defined herein line with OECD/PISA documentation (OECD, 2016), where teacher monitoring methods include several methods for monitoring teaching practices and teachers' work consisting of (a) tests or assessments of student achievement; (b) teacher peer review of lessons plans, assessment instruments, and lessons; (c) principal or senior staff observations of lessons; and (d) observation of classes by inspectors or other persons external to the school. Further, didaktik and curriculum education traditions are theoretically developed based on prior didaktik-curriculum dialogues involving Anglo-American and continental/Nordic scholars (Gundem & Hopmann, 1998; Westbury et al., 2000). Grouping of countries into respective didaktik and curriculum traditions follows prior categorization grounded on a rationale that is based on four criteria, namely historical, cultural, empirical, and practical (Tahirsylaj, 2019; Tahirsylaj et al., 2015).

Specifically, and to summarize the criteria for designating the 12 countries into respective didaktik and curriculum groupings, the historical criterion relates to the historical initiation and development of the didaktik tradition within German-speaking contexts in continental Europe, which then spread to the rest of continental and northern Europe, while the curriculum tradition emerged in the United Kingdom and then spread to other English-speaking countries (Kansanen, 1995). The cultural aspect is borrowed

---

[2] The article uses the original German term *Didaktik* throughout as already established in the literature (Westbury et al., 2000; Tahirsylaj, 2019) and to avoid the use of the English term *didactics*, which is often ascribed negative connotations such as frontal teaching (Kansanen, 1995).

from prior studies on world cultures, and more specifically the Global Leadership and Organizational Behaviour Effectiveness research project (GLOBE), which grouped world countries into ten cultural clusters based on data from surveys aimed at understanding organizational behaviour in respective societies (House, Hanges, Javidan, Dorfman, & Gupta, 2004), and the countries included in the present sample fall into Anglo-American, Germanic, and Nordic clusters accordingly, with Germanic and Nordic clusters forming the didaktik countries, while the Anglo-American cluster is represented by curriculum countries. The empirical criterion relies on empirical evidence from educational studies related to intraclass correlation coefficient (ICC) values across schools within countries, which show that the cultural clusters from world culture clusters referred to above could be a potential way to differentiate these clusters in terms of within-country school differentiation (Zhang et al., 2015). The practical element refers to the earlier didaktik-curriculum dialogue that took place during the 1990s when two groups of scholars were involved – scholars and researchers representing didaktik that included both German and Nordic scholars, and curriculum experts that included scholars mainly from the United Kingdom and the United States (Gundem & Hopmann, 1998). Interestingly, even when scholars seek to challenge the didaktik and curriculum categorization in recent work, mainly following a data-driven approach but not relying on any specific criteria, they build their studies following curriculum and didaktik grouping (e.g. Wermke & Prøitz, 2019). Nevertheless, the selection of countries in the two groups is purposeful in narrowing down the number of countries to the participating contributors in the initial didaktik-curriculum dialogue in the 1990s, and as such, it leaves out countries with a potentially didaktik tradition in places such as Iceland among the Nordic countries, the Netherlands or the Czech Republic among continental Europe countries.

A number of objectives guided the study. The first objective was to examine trends in teacher monitoring methods across schools within individual countries representing curriculum (primarily English speaking) and didaktik (continental and Nordic Europe) traditions. This is possible by the availability of representative large-scale PISA data sets. The second objective was to examine associations of TMM variables with students' reading, mathematics, and science performance within countries across the three PISA waves. The focus on the three different cognitive domains follows the PISA set-up, as in each cycle one domain is made the dominant one, and it corresponds with the data collected for TMMs.

Ultimately, the purpose of this cross-national study was not to simply compare and contrast countries but to discern the similarities and differences in national educational practices so that individual countries obtain a deeper understanding of their national educational orientations and practices. It has been argued that international comparative studies provide a rich set of understandings for national scholars, policymakers, and practitioners to help …

> […] *define what is achievable* […] *observe and characterize consequences of different practices and policies for different groups under different circumstances* […] *bring to light concepts for*

understanding education that have been overlooked […] *identify and question beliefs and assumptions that are taken for granted* (National Research Council [NRC], 2003, pp. 8–9, emphasis in the original)

To a considerable extent, this study is at the intersection of all these goals of international comparative studies as put forth by the NRC as it aimed to, first, theoretically explore similarities and differences between didaktik and curriculum orientations around TMMs, and second, to empirically examine how TMM practices are associated with students' performance in PISA across different national contexts. In this regard, the present study builds on the initial didaktik-curriculum dialogue[3] (see Westbury et al., 2000) by using TMMs from PISA data sets to empirically examine differences or similarities between the two groups of countries represented in the sample, following the four selection criteria listed above. The use of didaktik and curriculum perspectives to examine trends in TMMs is relevant since the goal of the article is to frame the study through educational lenses, rather than other approaches such as sociological, economic, and psychological ones that are routinely used to study educational phenomena. This framing is particularly relevant considering that it aligns with countries representing didaktik and curriculum traditions included in the sample.

Reviewing the types of international comparative studies in education, the NRC (2003) distinguished among Type I, Type II, and Type III international studies as per their primary purposes. The NRC (2003) defined Type I international comparative studies, primarily large-scale survey assessments, as ones that aim to compare educational outcomes cross-nationally and listed TIMSS and PISA as examples. Type II studies were designed to research specific educational policies and their implementation to inform educational policy in the United States. These studies were based on mixed-methods approaches, including quantitative, qualitative, descriptive, and interpretative studies and were primarily conducted on a lower scale, and studying high school-tracking in a couple of countries was listed as an example. Lastly, the NRC (2003) defined Type III studies as ones that aim to understand education more broadly, as well as to increase general understanding about education systems and processes. These studies could be either large- or small-scale and also mixed-methods oriented as Type II studies, and a study of culture and pedagogy in a set of five countries was listed as an example. Here too, the present study sits at the intersection of these three NRC-defined broad types of international comparative studies as it entailed gaining a better understanding of educational issues, such as TMMs, in a set of countries representing didaktik and curriculum educational systems (Type III), utilizing data from the PISA 2009, 2012, and 2015 waves (themselves Type I studies), to reach a better cross-national understanding

---

[3] The didaktik/curriculum dialogue in the 1990s identified a number of points of convergence and divergence between the two education traditions through a theoretical analysis, of which the most prominent were the primacy of content and its significance and Bildung in the didaktik tradition, and the emphasis on methods and evaluation in curriculum tradition. The present study is connected to this earlier international dialogue as it continues the comparison of the two traditions and their respective representative countries relying on empirical data and analysis.

and informing national policymakers in each national context of implications that the studied constructs might have for their school settings (Type II). To this end, the present study aimed to make a modest contribution to the ongoing and growing worldwide research efforts in the field of comparative and international education.

Next, the article turns to the conceptual framework and presents a brief literature review, followed by a general overview of education systems in didaktik and curriculum countries. Then, data and methodological considerations in data analyses are described, and findings and results presented, before ending with a discussion, conclusions, limitations, and potential avenues for further research.

Conceptual framework and literature review

For theoretical framing, the article draws first on didaktik/Bildung theory to place the discussion within continental and Nordic Europe educational thinking. Klette (2007) describes this tradition 'as a relation between teachers and learners (the who), subject matter (the what) and instructional methods (the how)' (p. 147). Hopmann (2007) states that didaktik is a matter of order, sequence, choice – features that align well with contemporary thinking regarding the United States' standards revision and attention to learning progressions (e.g. Duschl et al., 2011). Within the frame of order, sequence, and choice, Hopmann (2007) notes: 'Didaktik became the main tool for creating space for local teaching by providing interpretative tools for dealing with state guidelines on a local basis' (p. 113). Relatedly, the German concept of *Bildung* is a noun often translated in loose terms such as 'self-cultivation' or 'being educated, educatedness'. It also carries connotations of the word *bilden*, 'to form, to shape'. In Bildung, whatever is done or learned is done or learned to develop one's individuality, to unfold the capabilities of the I (Humboldt, 1793/2000). Overall, the didaktik tradition is defined as being more teacher-oriented, more content-focused, and one where there is a greater level of professional teacher autonomy (Deng & Luke, 2008; Tahirsylaj, 2019; Tahirsylaj et al., 2015; Westbury, 2000). Next, the article draws on the curriculum tradition, which, despite being categorized into several perspectives, such as humanistic, scholar, reconstructionist, and social efficiency, among others, has been dominated by the social efficiency orientation (Deng & Luke, 2008; Kliebard, 2004; Schiro, 2013; Schubert, 2008; Tahirsylaj, 2017; Tahirsylaj et al., 2015). Social efficiency promotes the preparation of future citizens with requisite skills, knowledge, and capital for economic and social productivity, while the subject matter is defined as practical or instrumental knowledge and skills that possess functional and utilitarian value (Deng & Luke, 2008). Overall, the curriculum tradition is defined as being more institution-oriented and teaching methods-focused, while also being more evaluation-intensive (Westbury, 2000). Therefore, it is relevant to address the issue of TMMs from the perspective of didaktik and curriculum traditions to examine the differences in the use of TMMs across the two groups of countries so that theoretical claims about the differences between the two traditions can be tested.

Then, the article turns to the transnational policy flows in education, where it has been argued that powerful knowledge producers and organizations such as the Organisation for Economic Co-operation and Development (OECD) and the World Bank, for example, influence national education policies in various forms – from curriculum designs to teacher education and accountability models, to name but a few (Drori et al., 2003; Sundberg & Wahlström, 2012). Relatedly, the scholarship on global education policy has found that education policies and practices have become more uniform across countries through a phenomenon known as 'institutional isomorphism' (Baker & LeTendre, 2005). From these perspectives, it is relevant to examine trends in teacher monitoring methods as defined by the OECD in PISA school questionnaires, because what gets measured has the potential to become mainstream in national education systems – especially in those that regularly participate in PISA, as is the case with the countries in the sample.

Next, the literature on teacher monitoring distinguishes between two purposes behind teacher monitoring and evaluation systems, namely summative and formative (Ford & Hewitt, 2020). Summative assessment of teachers relies on TMMs that are aimed at recognizing and/or rewarding teachers who perform better and punishing those that don't, while formative assessment is aimed at continuous professional development of teachers (Ford & Hewitt, 2020; Isoré, 2009). While it is not possible with the data available to distinguish the exact purposes of each of the TMMs within individual countries, it can be argued that the use of teacher peer reviews as a TMM could potentially be formative, while the use of students' tests or external inspectors could be linked to summative assessment of teachers' work.

The OECD PISA documentation shows that in 2015 all four TMMS were used across OECD countries in varying degrees:

> On average across OECD countries, 81% of students attend schools whose principals reported that tests or assessments of student achievement and principal or senior staff observations of lessons were used to monitor the practice of teachers; 66% attend schools that used teacher peer reviews of lesson plans, assessment instruments or lessons; and 42% attend schools where classes were observed by inspectors or other persons external to the school (OECD, 2016, p. 150).

The results shared by the OECD show that within-school monitoring of teaching practices is more predominant than external monitoring by inspectors or other persons external to the school. Nonetheless, the majority of students across OECD countries are now in schools where there is widespread monitoring of teaching practices and teachers' work.

The OECD findings have been confirmed by other independent academic research, which has found that teacher evaluation by school principals is one of the most widespread types of teacher evaluation worldwide (Brandt et al., 2007; Orphanos, 2014), but at the same time, the researchers have warned that teacher evaluation by school principals might be inflated or lack accuracy (Orphanos, 2014). The argument regarding school principals' inaccuracy might be further worsened in the PISA data since the data collected on TMMs are reported by school principals, and it can be claimed that if teachers themselves reported the data on how their teaching practices are monitored,

different results could be obtained from those reported by principals. This argument serves as a caution on how to understand and interpret the results shown later in the results section. Further, principals' lack of time exacerbates their role in teacher evaluation (Ridge & Lavigne, 2020). Next, a brief overview of education systems in sampled countries and data and method specifications is offered.

Context, data, and methods

A brief overview of education systems

All 12 education systems included in the study share similar structures in that they are expanding their education downward towards earlier early-childhood education, and then following with four or five years of elementary education, four or five years of lower secondary, and three or four years of upper secondary education. The exception is Ireland, which divides its education into two phases only – primary schooling covering ages 4–12, and post-primary schooling for ages 12–17/18. All countries have regulations for compulsory education in place, usually covering ages between 5/6/7 and 16. As regards the content covered in respective pre-university curricula, some differences between didaktik and curriculum countries are observed, most notably regarding the tracking of students along academic and vocational education paths, with curriculum countries providing a more comprehensive, less vocational-based education, and didaktik countries varying between less tracking in Nordic countries and more vocational-based tracking in Austria and Germany – with Germany being a clear outlier with the introduction of tracking as early as grade 5. Regarding monitoring and evaluation of schools' performance, countries have legislative frameworks in place that provide guidelines for internal and external evaluation of schools' and sometimes teachers' performance, and usually, this supervisory work is undertaken by national educational agencies focusing on pre-university education (e.g. *Skolverket* – the National Agency for Education – in Sweden) (Eurydice, 2019). In terms of PISA performance, all 12 countries have been performing and still perform either at or above the OECD average of 500 score points. The OECD calculates PISA test scores using item response theory (IRT) scaling to make possible descriptions of distributions of performance in a population or subpopulation being tested (OECD, 2017). PISA scores are presented then as plausible values, which are multiple imputations, with an average score of 500 and standard deviation of 100, and the scores represent degrees of proficiency in a particular domain (OECD, 2017). As Table 1 shows, only the United States in mathematics (470) and Austria in reading (485) achieve an average score that is statistically significantly below the OECD average in PISA 2015. Table 1 illustrates this point with results on three scales of reading, mathematics, and science in PISA 2015 across 12 countries in the sample, with didaktik and curriculum traditions represented by six countries each.

Table 1: Representative study sample of curriculum and didaktik countries and their test scores in PISA 2015 on three scales

Statistically significantly above the OECD average
Not statistically significantly different from the OECD average
Statistically significantly below the OECD average

	On the overall reading scale	On the mathematics scale	On the science scale
Finland	526	511	531
Canada	527	516	528
New Zealand	509	495	513
Australia	503	494	510
Norway	513	502	498
United States	497	470	496
Sweden	500	494	493
Germany	509	506	509
Ireland	521	504	503
Denmark	500	511	502
United Kingdom	498	492	509
Austria	485	497	495

Source: Adapted from the OECD, PISA 2015 database.
Note: Didaktik countries in red; curriculum countries in blue.

Data and methods

The study utilized data from PISA surveys administered in 2009, 2012, and 2015 and employed quantitative research methods. To address the first research question – How do curriculum and didaktik traditions compare across teacher monitoring methods using PISA 2009, 2012, and 2015 data? – a two-sample difference of proportion test to compare the means of TMMs for curriculum and didaktik samples was used. Because the subject domain focus rotates in PISA studies, the 2009 PISA data set collected data on TMMs focusing on methods used to monitor language teachers' practices, the 2012 data set on mathematics teachers' practices, and the 2015 data set on science teachers' practices. The descriptive procedure of the two-sample difference of proportion test was helpful to test the hypothesis that curriculum countries show greater use of the TMMs, namely tests, teachers, principals or inspectors, than didaktik countries. To address the second research question on the association of individual TMM variables with students' reading, mathematics, and science performance in PISA 2009, 2012, and 2015, respectively, inferential statistical analysis relying on hierarchical linear modelling (HLM) was utilized, to examine the effectiveness of TMMs in student performance and to capture the nested nature of PISA data (Raudenbush & Bryk, 2002). The second question is exploratory and no hypothesis was developed, mainly because no prior research has addressed the association of TMMs with student performance. The OECD initiated its work on developing PISA surveys around the mid-1990s and administered the first PISA survey in 2000 (OECD, 2002). PISA tests 15-year-old students' skills in three cognitive domains, namely mathematics, science, and reading. The countries representing didaktik

are Denmark, Finland, Norway, Sweden, Austria, and Germany, while the curriculum grouping comprises Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. Table 2 provides details about the sample size. The unequal sample size for the didaktik and curriculum groupings does not affect the results since to address the first research question through the two-sample difference of proportion test, the samples only need to be relatively large to obtain unbiased results and have to meet a threshold of at least more than five participants in both samples (Girdler-Brown & Dzikiti, 2018), which the present samples meet. The sample size doesn't affect results in the second research question either since the analyses are run separately for each country in the sample. The four criteria, namely historical, cultural, empirical, and practical, elaborated in the Introduction section, were applied to place countries into their respective didaktik and curriculum groupings.

Table 2: Representative study sample of countries, schools, and students from PISA 2009, 2012, and 2015

	2009		2012		2015	
Curriculum Countries	**Schools (N)**	**Students (N)**	**Schools (N)**	**Students (N)**	**Schools (N)**	**Students (N)**
Australia (AUS)	353	14251	191	4755	758	14530
Canada (CAN)	978	23207	230	5001	759	20058
United Kingdom (GBR)	482	12179	341	7481	550	14157
Ireland (IRL)	144	3937	311	8829	167	5741
New Zealand (NZL)	163	4643	197	4686	183	4520
United States (USA)	165	5233	209	4736	177	5712
Total	**2285**	**63450**	**1479**	**35488**	**2594**	**64718**
Didaktik Countries						
Austria (AUT)	281	6590	775	14481	269	7007
Germany (DEU)	226	4979	885	21544	256	6504
Denmark (DNK)	285	5924	507	12659	333	7161
Finland (FIN)	203	5810	183	5016	168	5882
Norway (NOR)	197	4660	177	4291	229	5456
Sweden (SWE)	189	4567	162	4978	202	5458
Total	**1381**	**32530**	**2689**	**62969**	**1457**	**37468**

Source: PISA/OECD database. The sample size across countries varies because countries decide how large their sample will be, and some countries, such as Australia, Canada, and the United Kingdom, test more students than is required for representativeness by the OECD/PISA.

Teacher monitoring methods (TMMs) data derived from four items included in the school survey completed by school principals. PISA started to collect data on TMMs for the first time in 2009. Principals responded to the following question: 'During "the last academic year", have any of the following methods been used to monitor the practice of teachers at your school?' The principals had the following options to choose from: (1) Tests or assessments of student achievement; (2) Teacher peer review (of lesson plans, assessment instruments, lessons); (3) Principal or senior staff observations of lessons; and (4) Observation of classes by inspectors or other persons external to the school. The question is phrased in the same way across all three PISA data waves used here. It should be recognized that data are reported by school principals only, and the information provided might have been different if teachers had reported it. To correspond with the focus of the PISA study in each wave, the TMMs in 2009 collected data related to language teachers, while in 2012 and 2015 the data related to mathematics and science teachers, respectively, and the trends across the four variables are presented in the results section. To address the second research question, students' reading, mathematics, and science performances in PISA 2009, 2012, and 2015, respectively, are used as dependent variables in the HLM models, while controlling for several independent variables (IVs), including student-level IVs such as socio-economic status (SES), gender, age, grade, immigration status and school-level IV: school type (public vs private). In the results section related to the second research question, only coefficients for the four key TMM items of interests in the full model are provided due to word limitations, and other detailed results are available upon request.

To develop the HLM models for PISA 2015 data, first, an unconditional model was run for each country using the dependent variable. The following is the specified equation for science achievement, while the same models apply for PISA 2009 and 2012 data, where science needs to be substituted with reading and mathematics scores, respectively.

$$science_{ij} = \beta_{0j} + e_{ij} \qquad\qquad (1)$$

Each school's intercept, β_{0j}, is then set equal to a grand mean, γ_{00}, and a random error u_{0j}.

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad\qquad (2)$$

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

Substituting (2) into (1) produces

$$science_{ij} = \gamma_{00} + u_{0j} + e_{ij} \qquad\qquad (3)$$

where:

β_{0j} = mean science achievement for school j

γ_{00} = grand mean for science achievement

Var (e_{ij}) = θ = within-school variance in science achievement

Var (u_{0j}) = $\square_{00}$ = between-school variance in science achievement

This model explains whether there is variation in students' 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 a linear mixed-effects model that splits the total residual or error into two error components. It starts with a multiple regression model, as follows:

$$\text{Science scores}_{ij} = \beta_1 + \beta_{2j} \, x_{2ij} + \ldots + \beta_p \, x_{pij} + \xi_{ij} \qquad (4)$$

Here β_1 is the constant for the model, while $\beta_{2j} \, x_{2ij}$ to $\beta_p \, x_{pij}$ represent covariates included in the given model. ξ_{ij} is the total residual that is split into two error components:

$$\xi_{ij} \; \Xi \; u_j + e_{ij} \qquad (5)$$

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 ξ_{ij} into the multiple linear regression model (4), we obtain the linear random intercept model with covariates:

$$\text{Science scores}_{ij} = \beta_1 + \beta_{2j} \, x_{2ij} + \ldots + \beta_p \, x_{pij} + u_j + e_{ij} \qquad (6)$$
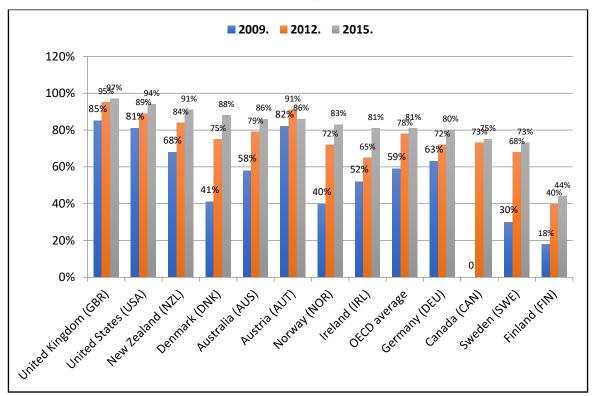
Again, $\beta_{2j} \, x_{2ij}$ to $\beta_p \, x_{pij}$ represent the covariates included in the model, and they vary depending on how many covariates are included in a specific model. The final model focuses on four level-2 covariates representing teacher monitoring methods whether teachers are monitored through student tests, teacher peer review, principal observation, or inspector observation, and it also includes one school-level covariate of school type (public vs private) as well as a number of student level-1 covariates including students' socio-economic status (SES), gender (girl), age, and immigration status, and controlling for dummy missing variables. A mean substitution was used to address missing data in HLM models. The same full HLM model was then run for each of the 12 countries in the study. The study relies entirely on secondary data analyses and does not create new scales of any sort, and as such relies on variables already in the PISA data sets; therefore, validity and reliability issues corresponding to TMM variables and reading, mathematics, and science test scores, for example, are dealt with in PISA documentation and technical reports (e.g. OECD, 2014).

Findings and results: the growing use of TMMs

In this section, results related to the first research question are presented first, followed by those of the second. In summary, the results indicate that within countries, the use of all four TMMs is becoming more widespread from one PISA wave to another irrespective of the domain the data come from, i.e. language, mathematics, or science teachers. In almost all 12 countries in the sample, the proportion of students in schools where any of the four TMM are used is lower in 2009 and increases progressively in 2012 and 2015. The increase is more dramatic for some of the countries, especially those under didaktik – such as Denmark and Norway – in the use of tests as a TMM than others (see Figures 1 to 4 for details on trends in the use of student tests, teacher peer review, principal
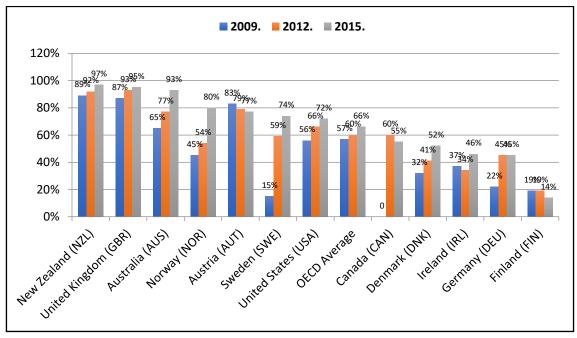
observation, and inspector observation as teacher monitoring methods within respective countries across three PISA waves, i.e. 2009, 2012, and 2015, where 2009 data cover school director responses on teaching practices of language teachers, 2012 to mathematics teachers, and 2015 to science teachers). Since the focus is on the trends in TMMs, the following results are shown per TMM across PISA waves rather than on subject-specific domains.

Figure 1: Proportion of students in schools where tests are used as a TMM per country in PISA 2009 (language teachers), 2012 (mathematics teachers), and 2015 (science teachers).
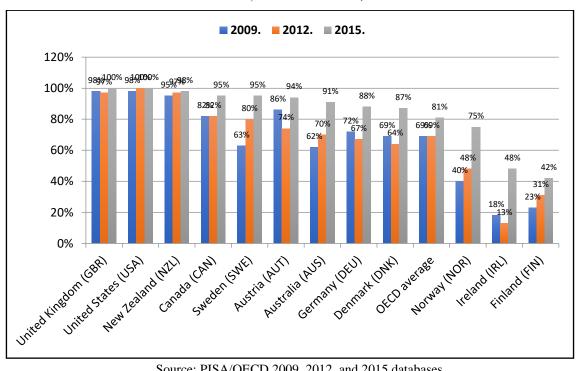

Source: PISA/OECD 2009, 2012, and 2015 databases

Figure 2: Proportion of students in schools where teacher peer review is used as a TMM per country in PISA 2009 (language teachers), 2012 (mathematics teachers), and 2015 (science teachers)
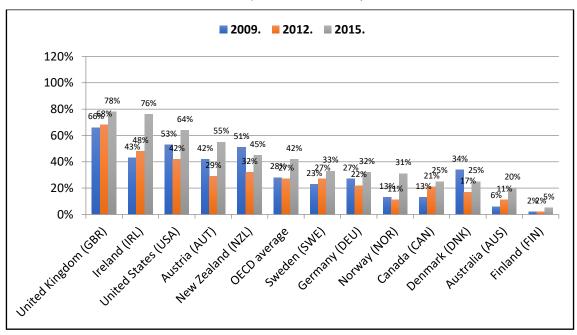

Source: PISA/OECD 2009, 2012, and 2015 databases

Figure 3: Proportion of students in schools where principal observation is used as a TMM per country in PISA 2009 (language teachers), 2012 (mathematics teachers), and 2015 (science teachers)

Source: PISA/OECD 2009, 2012, and 2015 databases

Figure 4: Proportion of students in schools where inspector observation is used as a TMM per country in PISA 2009 (language teachers), 2012 (mathematics teachers), and 2015 (science teachers)

Looking across countries, subject-specific domains, and PISA waves, and in each TMM specifically, the study finds that curriculum countries show greater use of all four TMMs than didaktik countries. This means that more students in curriculum countries are in schools where any of the TMMs is applied, indicating that teachers in the schools across curriculum countries are under more widespread monitoring than teachers in didaktik countries. It has also been observed that student tests, teacher peer review, and principal observation are much more common TMMs than external inspector observation across all countries. However, the use of external inspector observation is very low in several didaktik countries, and in the case of Finland almost nonexistent. The two-sample difference of proportion test showed that the differences in the use of TMMs in curriculum and didaktik countries are statistically significant and higher for curriculum than for didaktik countries in all four TMMs and across the three PISA waves of 2009, 2012, and 2015. The results validate the theoretical claims that curriculum countries are more evaluation-intensive both in terms of evaluating teachers through teacher monitoring and evaluating students through performance tests. While the observed trends in TMMs might also be partially attributed to the different teacher groups that data are associated within PISA 2009, 2012, and 2015, it is revealing that across countries the trends are consistent from 2009 to 2012 to 2015 and it is unlikely that each TMM was used more in 2015, for example, because the data are related to mathematics teachers in 2015. When more data become available through subsequent PISA data waves, it will be possible to test whether the same trends as observed here remain consistent since the data

will be related again to language, science, and mathematics teachers in the next three PISA waves, respectively. It is also worth noting that the trends do not reveal the frequency of use of any of the TMMs within and across countries as it is not known how many times a TMM has been used, however, the trends capture how widespread each of the TMMs is within schools within each country across the three PISA waves.

Associations of TMMs with students' reading, mathematics, and science performance in PISA 2009, 2012, and 2015 showed mixed estimates, with some being relatively large and significant for several countries, but they were not statistically significant for most of them as shown in Tables 3 to 5 below, which present detailed results from within-country HLM models with estimates for the four key variables of interest only. Results from the second research question to some extent point to the effectiveness of each TMM, i.e. does it matter whether one TMM is used over another for students' performance in PISA studies?

Table 3: HLM results only for TMM variables and their association with reading performance in PISA 2009 – curriculum and didaktik countries

	Australia	**Canada**	**United Kingdom**	**Ireland**	**New Zealand**	**United States**
Tests	12.58*	no data	/	/	/	/
Teachers	/	no data	/	/	/	/
Principals	/	-9.35*	21.71*	/	/	/
Inspectors	/	/	-20.52**	/	/	-33.68*
	Austria	**Germany**	**Denmark**	**Finland**	**Norway**	**Sweden**
Tests	/	/	/	/	/	/
Teachers	/	/	/	/	/	/
Principals	/	/	/	/	/	/
Inspectors	/	/	/	43.65***	/	/

Notes: "*" significant at p < 0.05, "**" significant at p < 0.01, and "***" significant at p < 0.001. "/" means estimates statistically not significant. Within-country HLM models control for students' socio-economic status (SES), gender (girl), age, immigration status, and school type (public).

The results from PISA 2009 in Table 3 indicate that TMMs were statistically significant in only four curriculum countries and one didaktik country out of 12 in the sample, and only in six cases – with the associations of *Tests* in Australia (12.58), *Principals* in the United Kingdom (21.71), and *Inspectors* in Finland (43.65) with reading test scores being relatively large and positive, while in three other cases coefficients were relatively large and negative for *Principals* (-9.35) in Canada, and *Inspectors* in the United Kingdom (-20.52), and in the United States (-33.68). Interestingly, in the United Kingdom, inspector observations are widely used as a TMM method, however, the negative coefficient is quite large and significant (-20.52); on the other hand, the use of inspector observation is associated positively with reading scores in the United Kingdom,

while in Finland inspector observations were reported to have been used as a TMM in schools covering only 2% of students; nevertheless, the coefficient is positive, relatively large, and significant (43.65).

Table 4: HLM results only for TMM variables and their association with mathematics performance in PISA 2012 – curriculum and didaktik countries

	Australia	Canada	United Kingdom	Ireland	New Zealand	United States
Tests	/	/	/	/	/	/
Teachers	/	/	-28.84*	/	/	/
Principals	/	/	/	/	/	38.35***
Inspectors	/	/	-36.49***	-13.62*	/	-22.07**
	Austria	**Germany**	**Denmark**	**Finland**	**Norway**	**Sweden**
Tests	39.79**	31.50*	/	/	/	/
Teachers	/	/	/	/	/	/
Principals	/	/	/	/	/	/
Inspectors	/	/	/	/	/	/

Notes: "*" significant at p < 0.05, "**" significant at p < 0.01, and "***" significant at p < 0.001. "/" means estimates statistically not significant. Within-country HLM models control for students' socio-economic status (SES), gender (girl), age, immigration status, and school type (public).

The associations of TMMs with mathematics scores in PISA 2012 data in Table 4 follow about the same pattern as in PISA 2009, being significant only in seven cases, although whenever significant the estimates were relatively large. In the United Kingdom, coefficients for *Teachers* (-28.84) and *Inspectors* (-36.49) were significant and negative, as they were for *Principals* in Ireland (-13.62) and the United States (-22.07), while coefficients for *Principals* in the United States (38.35) and *Tests* in Austria (39.79) and Germany (31.50) were relatively large, significant, and positive.

Table 5: HLM results only for TMM variables and their association with science performance in PISA 2015 – curriculum and didaktik countries

	Australia	Canada	United Kingdom	Ireland	New Zealand	United States
Tests	/	/	/	/	/	/
Teachers	/	/	/	/	/	/
Principals	-12.29**	/	/	/	-54.18***	/
Inspectors	/	/	/	/	/	/
	Austria	**Germany**	**Denmark**	**Finland**	**Norway**	**Sweden**
Tests	/	/	/	-16.61*	/	/

Teachers	/	/	/	/	/	/
Principals	/	/	/	/	/	/
Inspectors	18.49*	/	/	/	/	/

Notes: "*" significant at p < 0.05, "**" significant at p < 0.01, and "***" significant at p < 0.001. "/" means estimates statistically not significant. Within-country HLM models control for students' socio-economic status (SES), gender (girl), age, immigration status, and school type (public).

As Table 5 shows, the results were statistically significant only in four cases in PISA 2015 HLM models with science scores as a dependent variable – still negative in three cases in Australia (principals), New Zealand (principles), and Finland (tests), and positive only in Austria (inspectors). Overall, the evidence from HLM models suggests that different TMMs are associated differently with reading, mathematics, and science performance in PISA in various countries and across the three PISA waves, however, the associations had diminished by PISA 2015. This is a striking finding given that results from the first research question indicate that all four TMMs have been increasingly used with every PISA wave since 2009; however, the wider spread of each TMM does not seem to be associated with students' test scores in PISA 2015 as per the model fit and control variables included in the HLM models within each country and at least not consistently enough to determine a specific pattern. Nevertheless, in almost all cases when the results were significant, the estimates relatively firmly indicated a strong effect on PISA test scores.

Discussion and conclusions

The results of the study point towards a discernible convergence of practices in TMMs across countries as all TMMs are in use consistently but to varying degrees. However, the findings show statistically significant differences between didaktik and curriculum traditions in TMMs as in all three PISA waves and subject domains each of the four TMMs is used more in schools in the curriculum than in didaktik sample countries. However, some results might be over-reported: in previous survey work it was found that teacher self-assessment is the most common TMM in Sweden (Wahlström & Sundberg, 2017) and that parents are involved in teacher evaluations in Germany (Wermke & Prøitz, 2019), but neither teacher self-assessment nor parental assessment was included as options to be reported among TMMs in the PISA questionnaires. Nevertheless, as regards the first research question on the comparison across TMMs in didaktik and curriculum countries, the results are in line with the theory in that the curriculum tradition is more evaluation-intensive than the didaktik one, which is specifically promoted through the most dominant curriculum ideology of social efficiency (Deng & Luke, 2008; Tahirsylaj, 2017).

Next, even though didaktik and curriculum countries seem to belong to a continuum when examined individually, taken together as two distinct groups the results are in line with the theoretical divide found in the curriculum and didaktik scholarship. In this

regard, the results of the study challenge those of other studies, such as Wermke and Prøitz (2019), that seek to challenge didaktik-curriculum categorization. Here, it is shown that countries are predominantly aligned as expected within the didaktik and curriculum grouping in the four TMMs. Of course, there are a few exceptions, but the general trend is consistent and trends can only be observed in large samples at the country level – one school in Denmark and one school in Ireland, for example, only show idiosyncrasies for those specific school contexts and in no way are representative of the whole of Denmark and Ireland. By the same token, a school in Stockholm, the Swedish capital, and a school in Kiruna in northern Sweden might differ more than a school in the urban area of Stockholm and a school in the urban area of Vienna or Dublin.

Additionally, teacher peer review as a TMM could potentially be most relevant from the didaktik/Bildung tradition perspective to allow for opportunities for teachers to grow professionally as they go through the practice and thus serve as a form of formative assessment. However, teacher peer review was a more dominant TMM in the curriculum than didaktik countries overall, but still not positively and significantly associated with the students' PISA scores in any of the three waves in any of the countries. On the other hand, from the perspective of the social efficiency ideology of the curriculum tradition, inspector observation as a TMM could serve as a summative assessment of teachers' work, but as the results show, it is the least used TMM among the four across all countries, and in cases when it is significant for students' performance, it is consistently negative, particularly in the United Kingdom and the United States in PISA 2009 and 2012 results, in which countries' inspector observations are used more widely than in other countries in the sample. Only in Austria in PISA 2015 are inspector observations significant and positive for students' performance, thereby representing an outlier in the sample that could be a relevant case for exploration in further research.

Concerning the second research question, which tested whether TMM variables make a difference in students' reading, mathematics, and science performance as measured in PISA 2009, 2012, and 2015, respectively, the results are mixed but discouraging overall as only in a few cases were the TMM items significantly associated with students' performance in PISA, and far less in PISA 2015 than in the previous two PISA waves of 2009 and 2012. The lack of significant results may be primarily an effect of the global education reform movement in the Western world that has made education policy and practices more uniform across countries, a phenomenon known as 'institutional isomorphism' (Baker & LeTendre, 2005). The counterargument could also be made if the emphasis is placed on cases when TMMs do show strong associations with students' science performance, be it positive or negative, as is the case in four countries in PISA 2015. However, as Baker and LeTendre (2005) indicate, global similarities do not override national differences completely. Accordingly, findings here do highlight some national differences. Further, from the transnational policy flows in education, the findings indicate that specific measures promoted by international organizations such as the OECD through its PISA studies, including accountability through monitoring

teaching practices, have become widespread among all 12 countries in our sample from one PISA wave to the next, although more so in curriculum than didaktik ones. Thus, it is concluded that national education systems in didaktik countries still to some extent resist policies and measures promoted by the OECD. And by extension, the didaktik-curriculum education traditions continue to affect educational practices implemented in the given countries despite the external pressures, in the form of accountability measures, from supra-national and international organizations such as the European Commission in the context of Europe or the OECD (Tahirsylaj & Wahlström, 2019).

Interestingly, while TMMs are becoming more widespread across curriculum and didaktik countries from one PISA wave to the next their associations with students' performance in PISA are diminishing as shown by the results from the second research question. These revealing findings should serve as cautionary notes for education policymakers who are in favour of more teacher accountability measures since, as shown here, monitoring teachers' work and teaching practices are not consistently associated with improved student performance as per the data used and models fitted in the study. And considering the PISA 2015 results, the use of any TMMs can be questioned – why waste resources to monitor teachers' work through any of these monitoring methods when results are either widely insignificant or significantly negative and positive only in the case of inspectors in Austria? In sum, while accountability measures such as widespread control and intensive evaluation of teachers' work might serve other administration management purposes, they do not consistently and positively contribute to improved student performance and learning, and if only the latest PISA 2015 results are taken into consideration, such measures are statistically and substantively insignificant regarding students' performance.

Taken together, the findings of the study contribute to the field of comparative education pertinent to the recurring international didaktik and curriculum dialogue that was initiated in the 1990s. For example, an international project focusing on didaktik and curriculum studies has been taking place at the University of South Denmark (SDU), Denmark, since 2018. The project has been discussing, among other things, the relevance of the two education traditions for the present-day educational challenges. Specifically, then, the study's findings based on the use of recent comparable data offer an empirical understanding of similarities and differences in teacher monitoring methods across a set of representative didaktik and curriculum countries. As such, the study's empirical approach contributes to the diversification of methods and findings in the didaktik and curriculum studies field that has traditionally and historically been dominated by theoretical and conceptual research and scholarship.

Limitations and further research

The findings of this study may be generalized to specific countries that were examined more thoroughly here as well as to schools within them. However, there were several limitations, including the fact that the data used were cross-sectional and no causality is

intended with the findings. The study relied on secondary data analyses, therefore it was not possible to control what data were collected. The items used were derived from variables available in the PISA 2009, 2012, and 2015 data sets. The analytical models employed were specifically focused on associations of TMMs with students' reading, mathematics, and science performance, controlling for several variables; however, the variables used for predicting students' performance in PISA were selective and in no way exhaustive.

Teacher monitoring and evaluation are becoming increasingly more important with the expansion of policies aimed at using evidence for summative assessment of teacher performance in particular. To this end, it was important, first, to shed light on what TMMs countries have in place and what the trends are in their use across the years, and second, to examine the effectiveness of teaching TMMs in terms of their associations with student performance in PISA. The paper contributes towards increasing the understanding of and provides evidence about, trends in the use of TMMs in a set of 12 curriculum and didaktik countries. Further, the empirical evidence adds to the literature on the effectiveness of various TMMs on student performance, and also opens up further research opportunities to examine why and how some of the TMMs are more widely used in some countries than in others. Further research can specifically focus on how any of the TMM practices are implemented within specific school contexts and countries, and also further explore whether other TMM practices are used in addition to, or instead of, the four that PISA questionnaires have collected data on so far. Also, from the quantitative methodological perspective, utilizing quasi-experimental quantitative research methods such as propensity score matching (PSM) or regression discontinuity (RD) can potentially provide more precise estimates of TMM associations with students' performance in international large-scale assessments such as PISA. In sum, all 12 countries in the sample have been, and still are, relatively good performers in PISA studies with some fluctuation in scores across waves but mostly scoring at the OECD average or above. Therefore the issue will be about how effective they can be in the use of TMMs to maintain and/or improve high performance in PISA and the overall higher student achievement within countries that goes beyond PISA – to ensure that all students and teachers in the respective education systems have the opportunity to achieve their potential as learners and as professionals accordingly.

References

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

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). Examining district guidance to schools on teacher evaluation policies in the Midwest Region (Issues & Answers Report, REL 2007–No. 030). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. http://ies.ed.gov/ncee/edlabs

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational evaluation and policy analysis*, *24*(4), 305-331.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8-15.

Deng, Z. & Luke, A. (2008). Subject Matter: Defining and Theorizing School Subjects. In F. Connelly, M. F. He, & J. Phillion (Eds.), *The Sage Handbook of Curriculum and Instruction*. (pp. 66-87). Sage.

Drori G. S., Meyer, J.W., Ramirez, F.O., & Schofer, E. (2003). *Science in the Modern World Polity. Institutionalization and Globalization*. Stanford University Press.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*(2), 123-182.

Eurydice. (2019). *Welcome to Eurydice*. https://eacea.ec.europa.eu/national-policies/eurydice/home_en

Ford, T. G., & Hewitt, K. (2020). Better integrating summative and formative goals in the design of next generation teacher evaluation systems. *Education policy analysis archives*, *28*(63), 1-34. https://doi.org/10.14507/epaa.28.5024

Girdler-Brown, B. V., & Dzikiti, L. N. (2018). Hypothesis tests for the difference between two population proportions using Stata. *Southern African Journal of Public Health*, *2*(3), 63-68.

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

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73-112.

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

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

Humboldt, V. W. (1793/2000). Theory of Bildung. In I. Westbury, S. Hopmann, & K. Riquarts (Eds.), *Teaching as a reflective practice: The German Didaktik tradition* (pp. 57–61). Lawrence Erlbaum Associates.

Isoré, M. (2009). Teacher Evaluation: Current Practices in OECD Countries and a Literature Review, OECD Education Working Papers, No. 23, OECD Publishing. http://dx.doi.org/10.1787/223283631428

Kansanen, P. (1995). The Deutsche Didaktik. *Journal of Curriculum Studies*, *27*(4), 347-352.

Klette, K. (2007). Trends in Research on Teaching and Learning in Schools: didactics meets classroom studies. *European Educational Research Journal*, *6*(2), 147–160.

Kliebard, H. M. (2004). *The struggle for the American curriculum, 1893-1958*. Routledge.

National Research Council (NRC). (2003). *Understanding Others, Educating Ourselves: Getting More from International Comparative Studies in Education*. The National Academies Press.

OECD. (2017). *PISA 2015 technical report*. PISA/OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

OECD. (2016). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. PISA/OECD Publishing. http://dx.doi.org/10.1787/9789264267510-en

OECD. (2014). *PISA 2012 technical report*. PISA/OECD Publishing.

OECD. (2002). *PISA 2000 technical report.* PISA/OECD Publishing. https://www.oecd.org/pisa/data/33688233.pdf

Orphanos, S. (2014). What matters to principals when they evaluate teachers? Evidence from Cyprus. *Educational Management Administration & Leadership*, *42*(2), 243-258.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Ridge, B. L., & Lavigne, A. L. (2020). Improving instructional practice through peer observation and feedback: A review of the literature. *Education Policy Analysis Archives*, *28*(61).

Schiro, M. S. (2013). *Curriculum theory: Conflicting visions and enduring concerns*. Sage.

Schubert, W. H. (2008). Curriculum Inquiry. In Connelly, F. M., He, M. F., & Phillion, J. (Eds.). *The Sage Handbook of Curriculum and Instruction*. (pp. 399-419). Sage.

Sundberg, D., & Wahlström, N. (2012). Standards-based curricula in a denationalised conception of education: The case of Sweden. *European Educational Research Journal*, *11*(3), 342-356.

Tahirsylaj, A. (2019). Teacher autonomy and responsibility variation and association with student performance in Didaktik and curriculum traditions, *Journal of Curriculum Studies, 51*(2), 162-184. https://doi.org/10.1080/00220272.2018.1535667

Tahirsylaj, A., & Wahlström, N. (2019). Role of transnational and national education policies in realisation of critical thinking: the cases of Sweden and Kosovo. *The Curriculum Journal, 30*(4), 484-503. https://doi.org/10.1080/09585176.2019.1615523

Tahirsylaj, A. (2017). Curriculum Field in the Making: Influences That Led to Social Efficiency as Dominant Curriculum Ideology in Progressive Era in the U.S. *European Journal of Curriculum Studies, 4*(1), 618-628.

Tahirsylaj, A., Brezicha, K., & Ikoma, S. (2015), Unpacking Teacher Differences in Didaktik and Curriculum Traditions: Trends from TIMSS 2003, 2007, and 2011. In G. K. LeTendre & A. W. Wiseman (Eds.), *Promoting and Sustaining a Quality Teacher Workforce (International Perspectives on Education and Society, Volume 27)* (pp. 145-195). Emerald Group Publishing.

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.

Wahlström, N., & Sundberg, D. (2017). *Municipalities as actors in educational reforms: the implementation of curriculum reform Lgr 11*. Institute for Labor Market and Education Policy Evaluation, Uppsala, Sweden.

Wermke, W., & Prøitz, T. S. (2019). Discussing the curriculum-Didaktik dichotomy and comparative conceptualisations of the teaching profession. *Education Inquiry*, *10*(4), 300-327. https://doi.org/10.1080/20004508.2019.1618677

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). Lawrence Erlbaum Associates.

Westbury, I., Hopmann, S. & Riquarts, K. (Eds.). (2000). *Teaching as a reflective practice: The German Didaktik tradition*. 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. https://doi.org/10.1016/j.ijedudev.2015.02.013