Educational Assessment, Evaluation and Accountability (2019) 31:177–200
https://doi.org/10.1007/s11092-019-09297-5

# Teacher specialization and student perceived instructional quality: what are the relationships to student reading achievement?

**Stefan Johansson, et al.** *[full author details at the end of the article]*

Received: 20 July 2018 / Accepted: 15 April 2019 / Published online: 4 May 2019


**Abstract**
At an international level, teachers' work is increasingly circumscribed and regulated. Notions of accountability have shifted from primarily inputs to primary outcomes, and investment in strengthening teacher performance evaluation has expanded. At the same time, investment in enhancing the quality of teacher education programs is contested in many countries. Occupational professionalism, that is, a traditional, historic form characterized by discretionary decision-making, collegial authority, and trust in the practitioner, has been replaced by organizational professionalism that incorporates target-setting and performance review. The overarching question in this study concerns the meaningfulness and appropriateness of using student perceived instructional quality for the estimation of teaching quality in comparison to teacher specialization. The study investigates relations between fourth grade students' reading achievement levels, teacher specialization, and student perceptions of instructional quality, based on the Swedish PIRLS 2011 data. Performing two-level structural modeling with latent variables, this study revealed a positive relationship between teacher specialization relevant for the grade and subject taught, and student reading achievement. By contrast, there was no association between student perceptions of instructional quality and student reading achievement, or between instructional quality and teacher specialization. The results raise questions about the benefit of student evaluations of teacher classroom practices from both a validity perspective, as well as from a teacher professionalization perspective. However, the cross-sectional data used does not allow for causal inference, and further research on the relationships between teacher specialization, student perceived instructional quality, and student achievement is therefore needed.

**Keywords** Teacher education · Teacher specialization · Instructional quality · Structural equation modeling · PIRLS

## 1 Background

In recent decades, research has increasingly emphasized the importance of developing and maintaining a high-quality teaching force. There is now compelling evidence that

teachers account for a significant portion of variance in achievement between classrooms (Darling-Hammond 2000, 2014; Goldhaber and Anthony 2007; Hanushek 2011; Hanushek and Rivkin 2012; Hattie 2009; Hedges and Greenwald 1996; Kyriakides et al. 2009; Muijs et al. 2014; Nye et al. 2004; Rockoff 2004). Nonetheless, there are conflicting views on how best to guarantee teacher quality. While some researchers argue that investing in high-quality preparation is the most promising approach, others suggest that better facilitating entry to teacher education would attract stronger candidates (Boyd et al. 2009: Ingersoll 2007).

Issues about how best to design and organize teacher preparation have become increasingly prominent in undertakings to improve teacher quality. In this debate, there is an important underlying assumption that assessment of teacher quality in the classroom context can filter out poor-quality teachers, and effectively stimulate instructional improvement. However, in a review of existing literature, Hallinger et al. (2014) found little support for the belief that teacher evaluation represents a high-impact school improvement strategy.

In fact, a recurring theme in the instructional and school development literature emphasizes the potential costs and negative consequences of the frequent monitoring of teachers' work. Close surveillance and evaluation of teacher classroom behavior has been linked to teacher stress (Perryman et al. 2011), to de-professionalization (Zeichner 2010), to decreasing attractiveness of the teacher profession (Ingersoll et al. 2016), and to teacher attrition (Borman and Dowling 2008).

An important rationale for the scope of the current study is the increasing focus on efforts to categorize, scrutinize, and estimate teacher instructional behavior (Seidel and Shavelson 2007). The present study explores and compares relations between two measures of teacher quality—whereof one is a distal measure and the other is more proximal to teaching practice—and their relation to student achievement. Specifically, the study aims to investigate the relations between teacher specialization (defined as teacher education relevant for subject and grade), and student perceptions of instructional quality and fourth grade students' reading achievement.

## 2 Occupational and organizational teacher professionalism: opposing framings of teachers' work

In school systems that rely on accountability regimes, specialized teachers with deep subject knowledge and the ability to translate this knowledge into highly effective teaching find themselves under scrutiny. Teacher agency is under increasing threat from national governments, as statutory bodies are tasked with the control of curriculum content, and the assessment of teacher pedagogy and professionalism. Quite simply, reforms have changed what it means to be a teacher (Day 2007).

The concept of professionalization is widely used. It generally denotes an idea that certain occupations move towards strong occupational control (Abbott 1991). Professions are recognized as the mediators and applicators of knowledge in specific domains. The exclusive nature of skills and knowledge means that professionals have a mandate to make their own choices and decisions about proper interventions (Brante 2013). Professionalization has been described as the control of work, of scientific transformation, and the licensing of ethical conduct for the relation between

professionals and clients. The classical way of understanding professional work comprises a high level of trust (Lilja 2014).

In recent decades, schools and teachers in many countries have increasingly become subject to evaluation of varying kinds. Quality control and audit are regular elements. In education, professional work is both increasingly controlled and increasingly fragmented. It is characterized by external, low-trust accountability based on private sector strategies (Bottery 2006; Lilja 2014). In addition, the nature of audit culture demands quantification and measurement. Quality auditing builds on measures of performance that, in the main, eschew the use of peers or recourse to expert knowledge. Although schools are held accountable for the quality of their provision, it is often stakeholders beyond the teaching profession who determine what is to be prioritized. Consequently, the trust on which an expert system such as teaching depends is placed at risk (Perry 2006).

Abbott (1991) asserts that professions move in many directions, rather than in a single direction as implied by the term professionalization. Evetts (2006, 2011) distinguishes between occupational and organizational professionalism. Occupational professionalism, according to Evetts, is the more traditional, historical form and includes discretionary decision-making in complex cases, collegial authority, and the occupational control of work. It is based on trust in the practitioner by both clients and employers. Organizational professionalism, on the other hand, incorporates accountability, target-setting, and performance review.

Although national reforms differ in content, direction, and pace, there are important similarities. Day (2007) identifies three factors of particular relevance from a teacher quality perspective. First, reforms are proposed because governments believe that by intervening to change the conditions under which students learn, it is possible to accelerate improvements, raise standards of achievement, and thereby increase economic competitiveness. Second, the reforms result in an increased workload for teachers. Third, the reforms do not always pay attention to teacher identities. These identities are central to motivation, efficacy commitment, job satisfaction, and effectiveness. Day (2007) summarizes developments in England and Wales where schools are subjects to market pressures through parental choice of school, greater financial autonomy, evaluation, and target-setting. As outlined by Perryman et al. (2011), since the mid-1980s, teachers in England have lost a degree of relative autonomy, i.e., where they self-accountable though informal reflection and peer review. This relative autonomy has been replaced by loss of control, with stress being a common outcome. From an American perspective, Lavigne (2014) describes how, ever increasingly, hiring, firing, and tenure-granting policies are based on teacher evaluations. In a similar vein, Ingersoll (2005) explains how a teacher deficit perspective overlooks the organizational and occupational contexts of the work of teachers, which is characterized by low stature and social standing.

Much like developments in other countries, the influence that Swedish teachers have over their own work has diminished substantially. Teacher professionalism has been challenged or replaced by a teacher identity that no longer stresses the importance of specific and well-defined subject knowledge (Stenlås 2009). Teachers are accountable for perceived, as well as genuine shortcomings, and risk being "named, blamed, and shamed" (Dovemark and Holm 2017; Lindqvist et al. 2009). Occupational professionalism is contested by the frequent use of quality assessment from outside of the

profession. While trust in the quality of teacher education programs, and relevantly prepared teachers indicates a meritocratic view on teacher quality and occupational professionalism, reliance on close monitoring and the assessment of teacher classroom behavior suggests an organizational professionalism that is characterized by a loss of autonomy and trust.

## 3 The nature of effective teaching

Much research has been devoted to identifying actions and conditions that affect student outcomes (Seidel and Shavelson 2007). Though categorization varies across studies, a number of teaching process variables have been shown to positively affect student outcomes. Reinforcement, feedback, maximization of learning time, adaptive instruction, cooperative learning, and mastery learning are all examples of instructional strategies that have been found effective (Creemers and Kyriakides 2010; Darling-Hammond and Bransford 2005; Scheerens and Bosker 1997; Seidel and Shavelson 2007; Hattie 2009). It has also been argued that monitoring teacher behaviors is preferable, or indeed necessary, for guaranteeing positive student learning outcomes (Kyriakides et al. 2014). In the domain of literacy, Cunningham and Zibulsky (2009) noted that in the past three decades, the field of teacher knowledge has grown considerably, with studies specifically categorizing the knowledge and skills that teachers must acquire and apply. In this context, the development of instruments that can provide reliable and valid estimates of teacher knowledge has received considerable attention. Still, this work is based on the premise that it is possible to examine how the teacher knowledge base is associated with student outcomes, and as a consequence of this, to develop empirically validated best practices. However, solid evidence of the transformation from teacher knowledge into teacher practice is yet to be presented.

## 4 The role of teacher education for obtaining high-quality teaching

Despite the lack of compelling evidence of the nature and effects of transformation processes, there is an almost universal quest to improve teacher, and therefore also educational quality. With this comes the demand by policy-makers for higher quality teacher education (Imig and Imig 2006). According to Cochran-Smith et al. (2010), teacher education in the USA has begun to shift from preparing highly qualified teachers, to preparing highly effective teachers. This has resulted in a deprioritizing of university-based aspects of education, with an emphasis instead on practice. Notions of accountability have shifted from inputs to outcomes. As Darling-Hammond (2017) has noted, the knowledge base for teaching and the role of universities in preparing teachers is contested. In Australia and the USA, initiatives such as ‘Teach for Australia’ and ‘Teach for America’ involve the recruitment of candidates who enter teaching with just a few weeks of pre-service training. Cochran-Smith et al. (2015) suggest that there has been very little examination of alternatively certified teachers’ professional preparation in the USA. They are placed in classrooms largely based on the assumption that their previous knowledge, and tightly compressed preparation, is sufficient to equip them for teaching.

At the same time, there is growing evidence that expert knowledge is to a large degree attained through formal, high-quality teacher education during which content knowledge, pedagogical content knowledge, and general pedagogical knowledge is acquired (Adamson and Darling-Hammond 2012; Baumert et al. 2010; Croninger et al. 2007; Darling-Hammond et al. 2005; Depaepe et al. 2013; Kleickmann et al. 2013; Ball et al. 2008; Nye et al. 2004). Results indicate that teacher specialization with respect to subject and grade taught is important for effective teaching. Kunter et al. (2013) found that teacher characteristics that were not specific to the profession, such as for example general academic ability, had no relation to students' mathematics achievement or enjoyment. Instead, domain-specific knowledge had positive effects not only on student achievement, but also students' motivation. Simply being a smart student, they suggest, does not make somebody a good teacher. Further, effects of qualifications appropriate for grade level and subject can vary depending on subject domain and grade level (Wayne and Youngs 2003). For science and mathematics, and particularly at secondary level, a sizeable proportion of empirical results from Europe and the USA support the importance of specialized teachers (Baumert et al. 2010; Goe 2007). Wayne and Youngs (2003) argue that in the case of degrees, findings about the influence of coursework and certification have been inconclusive for all subjects other than mathematics. One reason for this result may be that mathematics is mostly learned in school, and that outcomes more sensitive to instruction than is the case for reading (Nye et al. 2004). While studies on the effects of teachers on reading at primary level are few (see, for example, Snow et al. 2005), evidence from the USA indicates an effect of relevant teacher education on reading achievement in lower grades (Clotfelter et al. 2007; Darling-Hammond 2014; Ferguson 1991; Nye et al. 2004). With data for Sweden, Johansson et al. (2015) estimated substantial effects of teacher education relevant for subject and grade on third graders' reading achievement levels, and found significant effects of teacher specialization on fourth graders reading achievement levels (Myrberg et al. 2018).

## 5 The role of student perceived instructional quality for obtaining high-quality teaching

Much endeavor has been dedicated to development of new models of teacher performance evaluation. "The dynamic model of educational effectiveness" was developed by Creemers and Kyriakides (2008), and has been used to measure teacher performance as perceived by students. Kyriakides et al. (2014) argue that secondary students are able to provide valid data on the classroom behavior of their teachers, and recently, studies have explored the validity of student ratings of instructional quality. The construct validity of ninth grade student ratings of instructional quality was investigated by Wagner et al. (2013). They found that while the structuring of teaching and classroom management could be generalized over classrooms and subjects (English and German languages), student motivation, the clarity of teaching, and the degree of student involvement could not. Similarly, Gaertner and Brunner (2018) investigated stability of student perceptions of instructional quality on class level over subjects, student grade levels, and for specific subjects. Results indicated that student ratings provided measures of

teaching constructs that were invariant across time, and for particular subjects, but not by grade levels. It was also suggested that young students may interpret certain item formulations differently than older students.

A few studies have related student ratings of instructional quality to student achievement. In a longitudinal study by Fauth et al. (2014), student ratings of teaching quality were related to science learning among third graders. While classroom management predicted achievement, supportive climate and cognitive activation did not. Here too, the researchers considered student ratings to be useful measures of teaching quality. Panayiotou et al. (2014) investigated relationships between concrete teacher actions in the classroom as reported by students, and student achievement gains in mathematics and science in six European countries. Though student prior achievement had by way the largest explanatory power, teacher behavior contributed with a small but significant part of the variation at student and class level. Results indicated that the student questionnaire was not equally interpreted between countries for all dimensions of teaching quality. The study design did not include data on teacher education.

In a comparative study of Nordic countries, and with large-scale data from TIMSS and PIRLS (mathematics, science, and reading) for fourth graders, Scherer and Gustafsson (2015) found that individual students tended to evaluate the teacher positively in the domains where they had performance strengths, and that student perceptions of how easy the teacher was to understand had significant relations to achievement in reading and maths between classrooms. On the other hand, Blömeke et al. (2016) noted substantial between-country differences in the relationship between student ratings of instructional quality, and fourth grade student mathematics achievement. Similarly, Nortvedt et al. (2016) investigated effects of the quality of teaching as measured by fourth grade students on reading achievement in 34 countries. They too found a largely inconsistent pattern, ranging from significant negative relationships to significant positive relationships. These researchers suggest that the varying sign and strength of the relationship between student assessments of instructional quality and achievement across countries is influenced by response styles and other, as yet unknown factors.

## 6 Summary of previous research

A traditional, occupational teacher professionalism, characterized by meritocracy, collegial authority, discretionary decision-making, and a high level of trust, is increasingly contested. In many countries, it is being replaced by an organizational professionalism that emphasizes accountability, target-setting, and performance review. External low-trust accountability is accompanied by a growing interest in reliance on student assessments of teaching quality.

Irrespective of the decreasing attractiveness of the teaching profession stemming from the frequent monitoring of teachers' work, it is important to note that there is little evidence supporting students' abilities to make accurate estimations of teacher quality (and which could potentially be related to student attainment). On the whole, results are inconsistent, and large between-country differences in the relation between student perceptions of instructional quality and student achievement levels are to be found. An overall inconsistency in results means

that it is wise to question the reliability of student assessments of teacher effectiveness. Furthermore, due to the large contextual differences between educational systems, country-specific analyses are warranted.

While there is a general agreement on the importance of teachers for student achievement, research on the effects of teacher education is still inconclusive. While some studies indicate that teacher education has little or no impact on teacher effectiveness, others have found it to be positively related to student outcomes. However, the relation between teacher education and student achievement is likely to be subject and grade specific. In particular, there is compelling evidence that teacher preparation is positively associated with student achievement in mathematics, and especially so in upper secondary grades. For teacher effects on reading, the picture remains unclear, though a growing body of research supports the importance of well-educated teachers. It has been suggested that more detailed and precise measures of teachers' education tend to be better predictors.

Few studies have investigated the relation between student perceptions of the quality of teaching, and formal teacher qualifications. Against a backdrop of contradictory opinions on the need for investments in maintaining and developing high-quality teacher education, there is a continuing need to shed light on the effects of teacher education on student achievement. The purpose of this study is therefore to investigate the relationship between teacher specialization, student assessed instructional quality, and student reading achievement. More precisely, the research questions are:

1) To what extent are teacher specialization and student reading achievement related?
2) To what extent are teacher specialization and student assessed instructional quality related?
3) To what extent are student assessed instructional quality and student reading achievement related?

## 7 Data and method

The empirical base for the study is Swedish data from the regularly recurring reading achievement study, Progress in International Reading Literacy Study (PIRLS), carried out by the International Association for the Evaluation of Educational Achievement (IEA). PIRLS assesses fourth graders' reading achievement in well over 50 countries. In the present study, we make use of Swedish data from the 2011 assessment, which contains a number of important add-ons to the international questionnaires. A national extension with additional background questions in the teacher questionnaire provides more detailed information on teachers' education. Students' assessments of the quality of teacher instruction were also expanded in the Swedish design. A representative sample of 4622 Swedish students and 218 teachers participated in the 2011 round of PIRLS.

### 7.1 Variables

The current study uses information from questionnaires from students, parents, and teachers. Student and teacher data was used to explore relationships between indicators of teacher specialization, student ratings of instructional quality, and student

achievement. Data from parents served as control variables. In the following sections, the variables used are presented, starting with the teacher variables.

### 7.1.1 Teachers

As previously mentioned, a national extension in the Swedish PIRLS 2011 teacher questionnaire provided unique data on aspects of teachers' education. Six education variables were considered for the current study: (1) type of teacher education, (2) emphasis on reading pedagogy, (3) preparation in teaching reading comprehension as a part of teacher education, (4) emphasis on Swedish language, (5) number of semesters studying Swedish language, and (6) focus on primary school-years during initial training. These six indicators define the latent variable "*TchSpec*." Together, they aim to capture specialization towards subject and grade. Cronbach's alpha for the six indicators (standardized) is .72. The variables are presented in the table below, along with descriptive statistics.

The variable indicating type of teacher education, "Tch_Ed," is based on an item comprising nine different teacher education programs. Mainly as a consequence of several teacher education reforms during recent decades, the type of teacher education varied substantially in the sample. Teacher education programs vary in the degree of relevance for teaching reading to fourth grade students. We therefore categorized the different education programs and recoded them into a dichotomous variable based on a categorization made in a previous study (Myrberg et al. 2018). The first category (code 1) comprises teachers with an education relevant for both subject and grade, that is, teaching reading in fourth grade. The other category (code 0) comprises all other teachers. For example, mathematics and science teachers, who held an education directed towards teaching in fourth grade, but were not educated for teaching reading, have been assigned the latter category.

The mean values of the indicators express the proportion for the dichotomized variables. As can be seen in Table 1, for a majority of teachers' education was aimed at the primary level. Further, a large proportion of teachers reported that they took courses with an emphasis on teaching methods. Also, many teachers answered that their education had a major emphasis on reading pedagogy and Swedish language. In addition to the teacher education variables, information provided by teachers on the total number of years of teaching experience, "TchExp," was used as a control in the analyses. In general, teachers were highly experienced, with on average over 16 years of teaching. While many students change teacher between the third and the fourth grade in Sweden, it is notable that more than half of the total number of teachers (54%) indicated that they taught the same class for two or more semesters. A total of 218 teachers responded to the questionnaire, whereof 84% were female. The average age was around 45 years and 34% of the teachers were in the age group 40–49. The response rates for the teacher competence items were some 90%.

### 7.1.2 Students

Six items from the student questionnaire, whereof one item was a national option, were used to operationalize the latent variable *student assessed instructional quality* (*Instr_qual*). A 4-point Likert scale ranging from "agree a lot" to "disagree a lot"

**Table 1** Descriptive statistics for teacher education variables

| Variable name | Label | *N* | Min | Max | M | SD |
|---|---|---|---|---|---|---|
| Tch_ed | Type of teacher education (specialized for grade and subject or not) | 201 | 0 | 1 | 0.53 | 0.50 |
| T_pedag | Emphasis on reading pedagogy (not at all - major emphasis) | 202 | 1 | 3 | 2.48 | 0.66 |
| T_did | Preparation in teaching reading comprehension as a part of teacher education (Y/N) | 194 | 0 | 1 | 0.89 | 0.31 |
| T_lang | Emphasis on Swedish language (not at all -major emphasis) | 204 | 1 | 3 | 2.77 | 0.51 |
| T_semest | Numbers of semesters studying Swedish language (0–4 semesters) | 180 | 1 | 6 | 3.57 | 1.55 |
| T_primary | Focus on primary school years (Y/N) | 212 | 0 | 1 | 0.90 | 0.30 |

was used for these items. Cronbach's alpha for the six items was .79. The variables are presented in Table 2.

A measure of student reading achievement was used as an outcome variable ("ReadAch"). The IEA provides five plausible values (PV's) for each individual's reading ability on a continuous scale (Martin et al. 2003). Based on Item Response Theory (IRT), reading achievement results for all students are placed on a common scale, even though they have not taken all the test-items. In IRT methodology, both individual and item attributes are taken into account when modeling a test result, this since the individual's achievement is considered to be a latent trait (e.g., Embretson and Reise 2000). IRT does not assume that the test scores include errors of measurement, only standard errors. As recommended by Rubin (1987), five separate analyses should be carried out—one for each PV. By averaging the results from the five runs, estimates are achieved. Since we performed two-level modeling, standard errors were pooled, taking into account both the between- and within-PV variances.

In order to account for stratification, case weights were used. To facilitate analyses with student and teacher data, IEA have provided weights for each hierarchical level, i.e., student and classroom. Weights at the classroom level are calculated by means of the product between a class weighting factor and a class weighting adjustment, as well as the product between a school weighting factor and a school weighting adjustment. At the student level, weights are a product of the student weighting factor and student

**Table 2** Descriptive statistics for variables indicating instructional quality as assessed by students

| Variable name | Label | *N* | Min | Max | M | SD |
|---|---|---|---|---|---|---|
| Intre_read | My teacher gives me interesting things to read | 4477 | 1 | 4 | 3.01 | 0.85 |
| Know_exp | I know what my teacher expects me to do | 4450 | 1 | 4 | 3.35 | 0.76 |
| Easy_und | My teacher is easy to understand | 4429 | 1 | 4 | 3.56 | 0.65 |
| Intre_tell | I am interested in what my teacher says | 4462 | 1 | 4 | 3.31 | 0.73 |
| Intre_task | My teacher gives me interesting things to do | 4468 | 1 | 4 | 3.17 | 0.75 |
| Expln_read (nat) | My teacher explains what I have to do to become a better reader | 4438 | 1 | 4 | 3.53 | 0.72 |

weighting adjustment (Foy 2013). In the present study, analyses were carried out using the multiple imputation option in Mplus 8 (Muthén and Muthén 1998-2018). This conveniently generates averaged results.

#### 7.1.3 Parents

Information on students' socioeconomic background (SES) was obtained from the home questionnaire completed by parents or guardians. Five indicators were used. One item indicates parents' estimation of the economic situation of the family in comparison with other families on a five-point scale, ranging from very well-off to not at all well-off. This item was a Swedish national option. Two items indicate the number of books in the household; the total number of books other than children's books on a five-point scale (ranging from 0 to 10 books to more than 200 books), and children's books only (on a five-point scale ranging from 0 to 10 books to more than 100 books). The item indicating mother's occupational status was reported on a five-point scale; remunerated work at least full-time, remunerated work part-time only, unremunerated work, other, and not applicable. It was decided not to use fathers' occupational status as the distribution in that variable was uneven and affected by ceiling effects, with 87% indicating working full-time. It should also be noted that the home questionnaire was mostly completed by mothers or female guardians—82 versus 33% for fathers (some had completed the questionnaire together). Finally, an indicator of mother's and father's education was used. The question was stated "What is the highest completed education by the mother and father respectively?" Seven alternatives (1–7), ranging from less than 9 years of compulsory schooling to post graduate education (PhD) were provided. Parents' education, "ParEd," was computed as a mean score of both parents' education. For example, if a student's mother held a PhD, and the father a bachelor degree, the estimated "score" is 6.5. Cronbach's alpha for the five SES indicators was .70.

#### 7.1.4 Further controls

Additionally, student language background was used as a control variable. This item related to how often Swedish was spoken at home. Three alternatives were given: "always or almost always," "sometimes," and "never." This item was coded 1–3, where 3 indicated "always or almost always." As shown in Table 2, most students often spoke Swedish at home.

PIRLS is a cross-sectional study that does not offer any pre-test of student achievement. However, parents have estimated their children's knowledge of written language before school start. This makes it possible to at least partly control for prior knowledge levels, and adds to the information on student background. By taking this information into account, the reliability of estimations of teacher effects increases.

To take students' early literacy skills into account, five indicators were retrieved from the parent questionnaire. On a four-point Likert scale (1 = Very well, 2 = Moderately well, 3 = Not very well, 4 = Not at all), parents estimated how well students could (1) recognize letters, (2) read words, (3) read sentences, (4) write letters, and (5) write words *at the time for school start*. Typically, this would be 3 years before the PIRLS assessment. Cronbach's alpha for this scale was .91 (Table 3).

**Table 3** Descriptive statistics of variables indicating student SES, language background, early literacy skills, and reading achievement

| Variable name | *N* | Min | Max | M | SD |
|---|---|---|---|---|---|
| Well-off finacially (well-off) | 3923 | 1 | 5 | 3.61 | 0.84 |
| Children's books (chbooks) | 3988 | 1 | 5 | 3.64 | 1.18 |
| Number of books (books) | 3523 | 1 | 5 | 3.68 | 1.23 |
| Mother's job (M_job) | 3869 | 1 | 5 | 4.32 | 0.99 |
| Parental education (ParEd) | 3653 | 1 | 7 | 4.17 | 1.36 |
| Recognize letters | 3994 | 1 | 4 | 3.42 | 0.69 |
| Read words | 3982 | 1 | 4 | 3.08 | 0.88 |
| Read sentences | 3973 | 1 | 4 | 2.46 | 1.00 |
| Write letters | 3982 | 1 | 4 | 3.18 | 0.75 |
| Write words | 3979 | 1 | 4 | 2.92 | 0.86 |
| Early Literacy (EarlyLit) | 4005 | 1 | 4 | 3.01 | 0.73 |
| Language background (Lang) | 4510 | 1 | 3 | 2.74 | 0.47 |
| Reading achievement (Read_ach) | 4622 | 265 | 784 | 539 | 67 |

### 7.2 Analytical approach

The main method of analysis is multilevel Structural Equation Modeling (SEM) with latent variables (e.g., Hox 2002). Compared to single indicator approaches, analyses that use latent variables have advantages in large samples, both theoretically and technically (Kline 2016). Since they are not directly observable, concepts within educational science are often difficult to frame with a single indicator. However, with latent variables, theoretical constructs can be represented in a more comprehensive manner. Moreover, latent variables are in a technical sense free from measurement errors. This is because the variance of different indicators not explained by the latent factor are sorted out. Thus, a latent variable comprises one "true" part (taking the common variance of the indicators into account), and one unexplained part (which is due to measurement error or score unreliability). From a validity point of view, this enables an analysis of better quality than would be the case in, for example, multiple regression analysis (MRA). In MRA it is assumed that all predictors are measured without error, which rarely is the case.

Educational assessment data often has a nested observational structure, e.g., students are clustered in classrooms. This means, for example, that students reading achievement scores within a classroom are not independent. The shared experiences/common influences among students within a classroom (same teachers, same classroom climate, peer-effects) are not repeated in any other classroom, and this dependence needs to be handled statistically. Assuming independence in analyses on hierarchical data would underestimate the standard errors. This in turn would lead to a too frequent rejection of the null hypothesis (Kline 2016). Hierarchically structured data is difficult to analyze; however, beginning in the 1980s, appropriate analytical methods have been developed through extensions of the basic regression model (e.g., Muthén 1989, 1991). Still, it was first in the

early twentieth century that computer programs with built-in capabilities for handling hierarchical data were developed. The analyses were conducted using SPSS version 24, and Mplus version 8 (Muthén and Muthén 1998-2018).

### 7.2.1 Model evaluation

For several decades, there has been a discussion around how to best assess model fit in SEM. One reason for the long-standing debate is that there are no golden rules or explicit cut-off values that indicate whether or not a model should be rejected or accepted (Bentler 2007; Fan and Sivo 2007; Goffin 2007; Markland 2007; Marsh et al. 2004). A reason for the lack of strict cut-off rules is that these values depend on factors such as the types of factor structures, sample sizes, and the size of factor loadings. In the current study, the $\chi^2$ goodness-of-fit test was used. Considering that the $\chi^2$ is sensitive to sample-size, it was combined with three other fit indices. RMSEA (Root Mean Square Error of Approximation) takes both the number of observations and free parameters into account. A RMSEA-value of 0.05 indicates a close fit, while a value of 0.08 has been suggested as acceptable (Loehlin 2004). The CFI (Comparative Fit Index) is a fit index that depends on the average size of the correlations in the data. CFI should be as close to 1.0 as possible, and 0.95 is considered as an acceptable value. SRMR (The Standardized Root Mean Square Residual), which is a measure of residual correlations computed separately for within and between levels, was also used. It has been suggested that the value of SRMR should be 0.08 or less for the model to be accepted (Brown 2006; Hu and Bentler 1999). When interpreting these cut-off values, it should be cautioned that these guidelines may not completely apply to multilevel SEM, as they have mainly been studied when SEM has been carried out using data at a single level.

### 7.2.2 Procedure

In the first step of the analysis, measurement models were formulated. Latent variables for teacher specialization (*TchSpec*) and student assessed instructional quality (*Instr_qual*) were formulated. As regards student socioeconomic background (SES) and student early literacy activities (EarlyLit*)*, mean scales were used. This was because latent measurement models did not have an acceptable fit.

The measurement models are presented in the "Results" section (Figs. 1 and 2). The latent teacher specialization variable was fitted at teacher level only. However, the other variables could be formulated at both student and teacher level (classroom means). Due to the many categorical items in the variable *TchSpec*, the WLSMV (robust weighted least squares) estimator was required. Hence, the model was fitted at one level only. This was because Mplus 8 does not allow WLSMV for two-level modeling. In a next step, and in order to facilitate two-level analyses with the teacher data, factor scores of *TchSpec* were saved and merged to the student level dataset. Since PIRLS data has a nested observational structure, with teachers being linked to groups of students, we could assign the factor scores of individual teachers to their students. The zero-order correlations for the variables used at the teacher level are presented in the table below (Table 4).

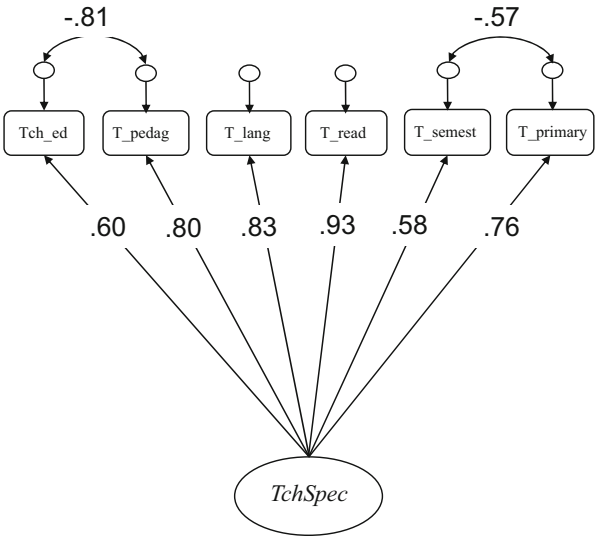


**Fig. 1** Measurement model of Teacher education. Factor loadings are all significant at $p < .01$

**Structural modeling** In a next step, we ran structural models according to a stepwise procedure. The first models seek to determine the relationships between *TchSpec*, *Instr_qual*, and student achievement. The purpose was to investigate whether the more specialized teachers taught with higher achieving classes, but also to see if students perceived the more relevantly educated teachers as providing better instructional quality. Thereafter, the explanatory variables, teacher experience, language background, SES, and early literacy abilities were introduced in the model one by one. Because the more specialized teachers may have been clustered together with groups of students with more advantageous backgrounds, we used SES and language background as controls. Furthermore, the students' early literacy abilities (Earlylit) was used as a proxy for students' prior achievement. As a next step, and in order to take into account any differential effects, we introduced a number of interactions.

**The rationale and procedure for testing interactions** A set of cross-level interactions were carried out in order to explore any possible compensatory effect of specialized teachers. In other words, we ran tests see if teachers with a more relevant education had a differentiated impact with respect to student-SES. An interaction between achievement and student assessed instructional quality was also tested to investigate whether high or low student achievement levels had a relation to students' estimations of the quality of instruction. In such an interaction, the slopes and intercepts are assumed to vary across classrooms, and are thus specified as random latent variables at the teacher level. These specifications correspond to cross-level interactions where the regression of achievement on SES at the student level varies as a function of teacher competence at the between-level part of the model. The results from these tests are presented under the heading "Structural models" in the "Results" section.

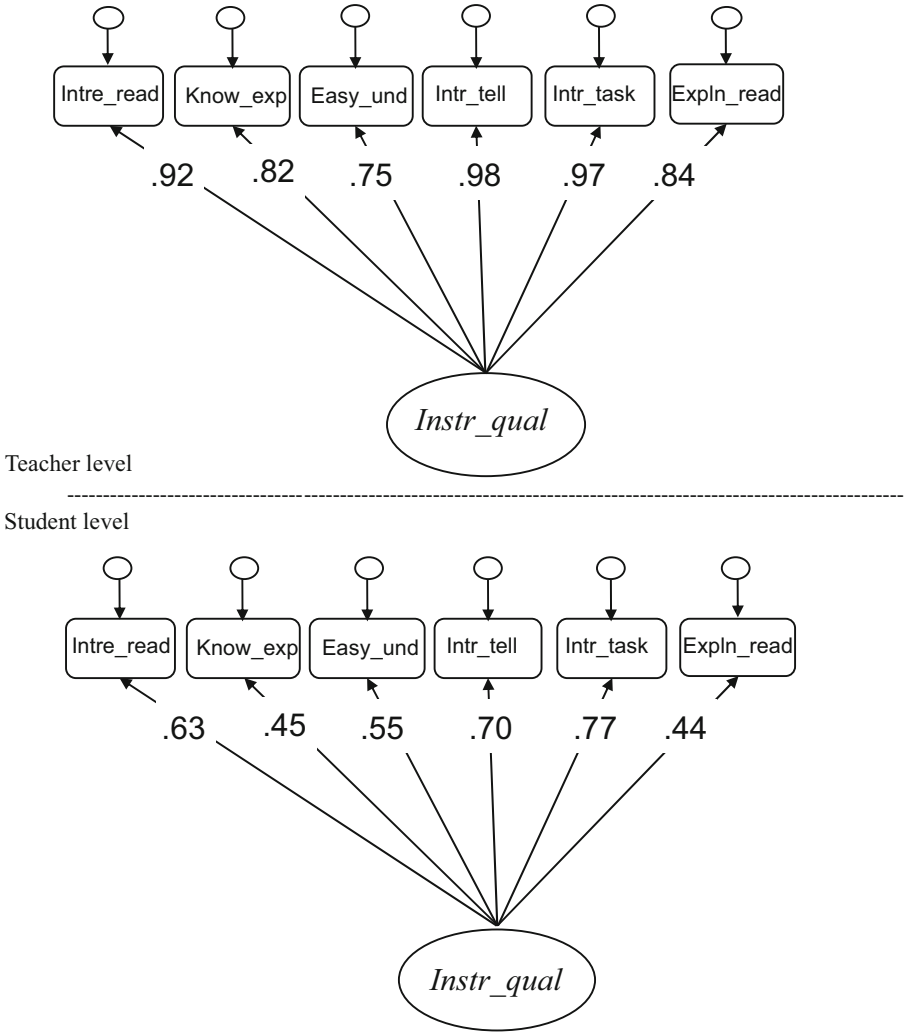


**Fig. 2** Measurement model of student assessed instructional quality. All factor loadings are significant at $p < .01$

## 8 Results

The first step in the modeling procedure was to fit the two latent measurement models to the data. In Fig. 1, the teacher education variable (*TchSpec*) is presented with its factor loadings.

Six indicators formulated the latent variable *TchSpec*. A covariance was included between the residuals of "T_pedag" and "Tch_ed," and "T_primary" and "T_semest." The covariance between the error terms was negative, which shows that they share unique variance which is not absorbed by the latent factor, where high values in one of the variables correspond to low values in the other variable. The negative covariance

**Table 4** Zero-order correlations for variables at between level (teacher level)

| Variables | TchSpec | T_exp | Ach | Intr read | Know_exp | Easy_und | Intr_tell | Intr _task | Expln_read | M_job | Books | Chbook | Par_ed | Well-off | Lang |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TchSpec | 1.00 | | | | | | | | | | | | | | |
| T exp | 0.17 | 1.00 | | | | | | | | | | | | | |
| Read_ach | 0.19 | 0.05 | 1.00 | | | | | | | | | | | | |
| Intr_read | − 0.05 | 0.29 | − 0.13 | 1.00 | | | | | | | | | | | |
| Know_exp | 0.04 | 0.27 | 0.04 | 0.69 | 1.00 | | | | | | | | | | |
| Easy_und | − 0.00 | 0.24 | 0.09 | 0.78 | 0.66 | 1.00 | | | | | | | | | |
| Intr_tell | − 0.05 | 0.20 | − 0.08 | 0.82 | 0.84 | 0.77 | 1.00 | | | | | | | | |
| Intr_task | − 0.04 | 0.16 | − 0.09 | 0.85 | 0.76 | 0.74 | 0.97 | 1.00 | | | | | | | |
| Expln_read | 0.21 | 0.29 | − 0.27 | 0.82 | 0.63 | 0.53 | 0.75 | 0.80 | 1.00 | | | | | | |
| M_job | 0.17 | 0.05 | 0.70 | − 0.09 | − 0.08 | 0.10 | − 0.10 | − 0.03 | − 0.19 | 1.00 | | | | | |
| Books | 0.08 | 0.07 | 0.85 | − 0.19 | − 0.09 | 0.11 | − 0.11 | − 0.12 | − 0.39 | 0.73 | 1.00 | | | | |
| Chbooks | 0.96 | 0.08 | 0.07 | 0.85 | − 0.19 | − 0.09 | 0.11 | − 0.11 | − 0.12 | − 0.39 | 0.73 | 1.00 | | | |
| ParEd | 0.02 | 0.04 | 0.69 | − 0.03 | 0.15 | 0.19 | 0.16 | 0.11 | − 0.25 | 0.59 | 0.82 | 0.82 | 1.00 | | |
| Well-off | 0.05 | 0.03 | 0.40 | 0.23 | 0.42 | 0.28 | 0.45 | 0.37 | 0.10 | 0.46 | 0.50 | 0.50 | 0.79 | 1.00 | |
| Lang | 0.07 | 0.00 | 0.53 | − 0.20 | − 0.05 | 0.16 | − 0.20 | −0.16 | − 0.19 | 0.62 | 0.63 | 0.63 | 0.23 | 0.09 | 1.00 |

between the indicators "T_primary" and "T_semest" is reasonable, as teacher education programs for primary teachers are usually shorter than programs directed towards middle school. Nevertheless, since both these variables were considered important for reasons of construct validity, they were kept in the model of *TchSpec*. The model obtained good fit to the data. Fit statistics for all the presented models are displayed in Table 3. The factor loadings for the latent variable *TchSpec* are all significant, and are moderate to high. Factor scores were saved for the latent variable and merged onto the student level data set, thus creating a continuous variable used in the further analyses applying two-level modeling.

In the next step, student assessed instructional quality was modeled using seven indicators from the student questionnaire. The measurement model is presented in Fig. 2.

The measurement model of student assessed instructional quality "*Instr_qual*" could be formulated at two levels. At teacher level, the classroom averages for the items are estimated. The factor loadings are all significant and fairly high, especially so at the between level. Fit statistics for the latent variables are presented in Table 5.

## 9 Structural models

In the next step, we investigated if (1) teacher specialization (*TchSpec*) was related to student reading achievement levels, if (2) teacher specialization was related to student assessed instructional quality, and if (3) student assessed instructional quality was related to student reading achievement levels. In the first model (model 1), a positive relation between teacher specialization and student achievement of .19 was found. In model 2, student assessed instructional quality was introduced both as a dependent and an independent variable. Unexpectedly, student assessed instructional quality was uncorrelated with teacher specialization. Further, no significant relation was found between instructional quality and student reading achievement. Consequently, students (or classrooms) with higher achievement did not rate their teachers higher. In model 3, teaching experience (T_exp) was introduced as an explanatory variable. While the previously estimated coefficients remained about the same, "T_exp" had a positive relation to *Instr_qual* of .21. This indicates that the students taught by more experienced teachers perceived them as providing better instructional quality. Notably, with the effect of teacher specialization under control, the relation between experience and student achievement was zero. The correlation between "T_exp" and *TchSpec* was modest, albeit statistically significant, amounting to .18. In model 4, students' early literacy abilities were accounted for. While "EarlyLit" had a substantial positive relation

**Table 5** Model fit for the two measurement models of teaching quality

| Fit statistics | Chi2 | Df | p | RMSEA | CFI | SRMRw | SRMRb |
|---|---|---|---|---|---|---|---|
| Model | | | | | | | |
| *TchSpec* | 11.922 | 7 | .103 | 0.057 | 0.988 | | |
| *Instr_qual* | 100.170 | 18 | .000 | 0.032 | 0.979 | 0.025 | 0.045 |

with student achievement, it did not have any influence on the relation between teacher specialization and achievement. In model 5, SES was taken into account, mainly as a control for potential selection effects (well-educated teachers clustered together with students from more advantageous background). When SES is introduced in models linking measures of teacher competence to student achievement, it can often be anticipated that the SES-effects overshadows any other effects. Interestingly, in this model, the effect of *TchSpec* on "Read ach" was reduced, but nevertheless remained significant at $p < .10$. The relation between "T_exp" and *Instr_qual* did not change when "SES" was accounted for. In model 6, we included "Lang" in the model. However, language background seemed to be confounded with both "EarlyLit" and "SES," since no significant effect could be observed for any of these three variables. Therefore, in order to avoid multicollinearity, and to shed more light on the relationship between "Read ach," "SES," and "Lang," in model 7, "EarlyLit" was deleted from the model. It could be noted that "Lang" did not have an influence which went beyond "SES." In line with previous evidence, "SES" had a strong relation to achievement. Results reported for the teacher level are presented in Table 6.

In order to test the robustness of results, we used two modeling approaches. First, we ran ordinary regression analysis using cross-products on an aggregated teacher level data set. However, multicollinearity was observed, and the accuracy of estimates was not deemed trustworthy. Then, we tested a set of interactions in a multilevel model (see,

**Table 6** Relationships between teacher qualification and reading achievement presented at the teacher level

| Dependent | Predictor | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Read ach | | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
| | *TchSpec* | .19* (.08) | .19* (.08) | .18* (.09) | .18* (.09) | .11 (.07) | .14 (.08) | .11 (.07) |
| | Instr_qual | | − .09 (.13) | − .10 (.14) | − .21 (.15) | − .10 (.10) | − .10 (.10) | − .06 (.09) |
| | T_exp | | | .03 (.09) | .04 (.09) | .01 (.06) | .01 (.06) | .01 (.06) |
| | EarlyLit | | | | .47* (.12) | .08 (.14) | .53 (.33) | |
| | SES | | | | | .75* (.09) | .26 (.34) | .73* (.09) |
| | Lang | | | | | | .56 (.32) | .11 (.08) |
| *Instr_qual* | | | | | | | | |
| | *TchSpec* | | − .02 (.13) | − .05 (.13) | − .05 (.13) | − .05 (.13) | − .05 (.13) | − .05 (.13) |
| | T_exp | | | .26* (.10) | .26* (.10) | .26* (.10) | .25* (.10) | .25* (.10) |
| | EarlyLit | | | | .22 (.14) | .23 (.14) | .08 (.14) | |
| | SES | | | | | .00 (.11) | .09 (.15) | .10 (.14) |
| | Lang | | | | | | − .20 (.15) | − .20 (.15) |
| Model fit | | | | | | | | |
| Chi2/Df | | | 163.9/33 | 171.7/38 | 191.5/48 | 285.9/59 | 251.23/68 | 232.2/58 |
| RMSEA | | | .029 | .028 | .025 | .029 | .024 | .025 |
| CFI | | | .968 | .968 | .969 | .956 | .965 | .964 |
| SRMRw | | | .027 | .027 | .026 | .036 | .026 | .027 |
| SRMRb | | | .071 | .068 | .073 | .079 | .08 | .08 |

Standardized coefficients and their standard errors (in parentheses). * Coefficients are significant at $p < .05$

Hox 2002). Here, random slopes were specified for the following relations: Read_ach on Lang; Read_ach on SES; Read_ach on EarlyLit; Read_ach on *Instr_qual*. Thereafter, the between-level variables *TchSpec* and T_exp were used to investigate possible interaction effects. However, there was no significant variation in the slopes, and no changes occurred when we related the teacher variables to the slopes. This indicates that relations did not vary between classrooms. For example, the effect of SES on reading achievement was similar across classrooms. It should though be noted that the number of clusters in this study might have been too limited to identify differential effects with the use of the random slope technique.

In summary, the main result from the relationships tested in a series of structural models is that teacher specialization was linked to Swedish fourth grade students' reading achievement levels, while student perceptions of instructional quality were not.

## 10 Discussion and conclusion

This study explored and contrasted two measures of teacher quality and their respective relations with fourth grade students' reading achievement levels. A Swedish national extension made possible the use of a number of indicators of teacher specialization with respect to preparation for the subject and the grade taught. A positive, significant relation between teacher specialization and student achievement was observed, which is in accordance with both previous results from Sweden (Myrberg 2007; Johansson et al. 2015, Myrberg et al. 2018), and a substantial amount of international research (e.g., Adamson and Darling-Hammond 2012; Kleickmann et al. 2013; Nye et al. 2004).

Teacher specialization is a distal measure of teacher quality. Student perception of instructional quality is however a proximal measure, and one increasingly used to evaluate teacher performance. It has been suggested that young students are well placed to evaluate the qualities of teaching, and particularly so with regard to aspects of classroom management (for example, management of time and disorderly student behavior) (Fauth et al. 2014; Kyriakides et al. 2014; Panayiotou et al. 2014). We used student questionnaire data covering a range of aspects of instructional quality, and related it to student achievement. However, contradictory to some previous research, no association between student perceptions of instructional quality and achievement was detected. It should, though, be noted that student perceived instructional quality as measured in the present study was intended to capture cognitive activation, academic focus, and clarity of instruction, rather than aspects of classroom management and student misbehavior.

Comparative studies have concluded that, as regards the relationship between students' perceptions of instructional quality and achievement, results are generally inconclusive (Blömeke et al. 2016; Nortvedt et al. 2016). In particular, current knowledge on the predictive power of primary school students' perceptions is limited. As pointed out by Nortvedt et al. (2016), a highly inconsistent pattern can be observed between countries in the association between fourth graders' reading achievement and their perceptions of instructional quality, with the causes of between-country differences poorly understood.

The results of the current study support the idea that while teacher specialization can be linked to effective teaching practices affecting student achievement, students might not be able to recognize important aspects of instructional quality influencing achievement levels.

### 10.1 Implications

Our results have implications for practice, policy, and research beyond the Swedish context. As Hallinger et al. (2014) have noted, the broadening consensus on the importance of teaching quality has emerged during an era of expanding educational accountability. It is an era too where a number of negative effects emanating from the frequent measurement and surveillance of teacher classroom behavior, such as for example, stress and decreased job satisfaction, are also recognized (Zeichner 2010).

It has previously been proposed that proximal measures of effective teaching practices, such as student perceptions of instructional quality in classrooms, are preferable, or indeed necessary, in studies of instructional effectiveness (Seidel and Shavelson 2007). Kyriakides et al. (2014) have for example suggested that student questionnaire data can be used to identify individual teachers' professional needs, and in guiding the development of area-specific courses for each teacher and school. In this was data can be used for individual school improvement.

However, the absence of an association between student perceptions of instructional quality and achievement in this study raises questions about the meaningfulness or appropriateness of using (younger) students' accounts of teaching quality to evaluate teacher effectiveness. Instead, we would like to point to the potential value of teacher specialization, a measure more distal to classroom practice. The basis of this argument is twofold. First, a growing body of research demonstrates substantial, positive effects of teacher education in general, and teacher specialization in particular, on student achievement. Second, the close surveillance and constant evaluation that, in many countries, teachers experience, is associated with distrust and de-professionalization. A working environment where student evaluations are used to measure teacher quality enhances the risk for teachers adopting strategies intended to satisfy student opinions, rather than being based on long-term, didactical deliberation (Perryman et al. 2011; Dovemark and Holm 2017). In order not to further damage the attractiveness of the teaching profession, and to not further erode teacher professionalization, distal measures of teacher quality may be preferable if basic reliability criteria are met. If teacher quality can be assured by employment of teachers who are appropriately educated for the job, it would positively affect the possibilities for teachers to formulate and further develop quality standards in collegial collaboration. Development of effective teaching practices could thus be redirected to the teacher community, and the (external) influence of other stakeholders could be restricted. This would probably strengthen professional claims, and likely increase the attractiveness of the teacher profession, as well as the quality of education (Borman and Dowling 2008; Cochran-Smith et al. 2010; Ingersoll et al. 2016).

Bottery (2006) highlights a need for education professionals to add their voice in a larger dialogic project. We would like to add that in the complex professional work of teaching, it is not possible or even desirable to make all the theoretical and empirical considerations underpinning actions visible, transparent, and auditable. Expert teachers use a wide variety of approaches, tools, and methods that neither can, nor should be totally transparent or entirely understandable to others. Instead, well-educated and specialized teaching professional should be given the opportunity to develop successful teaching practices collegially, and in forms characterized by autonomy and trust.

### 10.2 Limitations and suggestions for future research

The classroom was the primary level of interest in this study, and two-level analysis was used. This approach takes account of the variation in individual student background characteristics likely to interfere with teacher effects. Ideally, the organization of students within classes within schools should be accounted for by means of three-level analysis. However, the sampling procedure of PIRLS in Sweden does not allow for such analyses, as only some 20 of the schools included in the study participated with more than one classroom. Restriction of analysis to two levels could have affected results. This is because the aggregated knowledge-level of the teachers in a particular school can be anticipated to exert an influence that extends beyond the individual classroom. This having been said, a substantial body of previous research suggests that the variation between classrooms is likely to be far more significant than the variation between schools (Hill and Rowe 1996; Luyten et al. 2005).

When interpreting results, it should be considered that teacher effects are likely to vary across grades and subjects. Luyten (2003) found support for larger teacher effects in primary school than in secondary school. Also, previous research indicates that mathematics seems more sensitive to classroom instruction than reading, probably because reading skills are more likely to also be acquired outside of school (Clotfelter et al. 2007; Goe 2007; Nye et al. 2004). Teacher effectiveness is most certainly subject- and grade-specific; it might also be differentiated with respect to student characteristics. There is still a relatively small number of studies that has investigated teacher effects on achievement. Consequently, further research on this relationship would be of significant value.

The cross-sectional design of PIRLS does not allow for causal inferences to be made, as prior knowledge cannot be accounted for in the estimation of the strength of relationships between independent and dependent variables. Nevertheless, we were able to control for student language and social background, accounting for effects of social selection. In a further attempt to control for student prior knowledge, information provided by parents on students' language skills when formal schooling started was included. A significant portion of variance could thus be sorted out this way, thus strengthening the validity of estimated effects.

Ideally, studies investigating teacher effects would employ rigorous designs. As the resources needed for experimental and longitudinal designs aimed at studying the effects of teacher education are restricted, the possibilities attaching to large-scale international comparative studies should be highlighted. Especially at country level, school systems can to large extent function as their own controls, as many contextual factors within school systems do no change substantially over time. Questionnaires could therefore be further developed to gather more precise and detailed measures of teacher education, its length, and content. We believe that research would benefit from this type of development. It is imperative that increased knowledge is generated about these important features of effective teachers, and the role that teacher specialization has for student achievement.

**Funding** This research was supported by grants from the Swedish Research Council (grant number 721-2013-2207).

## References


Abbott, A. (1991). The order of professionalization. An empirical analysis. *Work and Occupations, 18*(4), 355–384.

Adamson, F., & Darling-Hammond, L. (2012). Funding disparities and the inequitable distribution of teachers: evaluating sources and solutions. *Education Policy Archives, 20*(37), 1–42. https://doi.org/10.14507/epaa.v20n37.2012.

Ball, D. L., Hoover Thames, M., & Phelps, G. (2008). Content knowledge for teaching. What makes it special? *Journal of Teacher Education, 59*(5), 389–407. https://doi.org/10.1177/0022487108324554.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and pupil progress. American Educational Research Journal, 47, 133–180. doi:https://doi.org/10.3102/0002831209345157.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*, 825–829. https://doi.org/10.1016/j.paid.2006.09.024.

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes* (pp. 21–50). IEA Research for Education (A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA)), vol 2. Springer, Cham.

Borman, G. D., & Dowling, N. M. (2008). Teacher attrition and retention: a meta-analytic and narrative review of the research. *Review of Educational Research, 78*(3), 367–409. https://doi.org/10.3102/0034654308321455.

Bottery, M. (2006). Education and globalization: redefining the role of the educational professional. *Educational Review, 58*(1), 95–113. https://doi.org/10.1080/00131910500352804.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Eductional Evaluation and Policy Analysis, 31*(4), 416–440. https://doi.org/10.3102/0162373709353129.

Brante, T. (2013). The professional landscape: the historical development of professions in Sweden. *Professions and Professionalism, 3*(2), 1–18. https://doi.org/10.7577/pp.558.

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

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: longitudinal analysis with student fixed effects. *Economics of Education Review, 26*, 673–682. https://doi.org/10.3386/w12828.

Cochran-Smith, M., Cannady, M., McEachern, K. P., Piazza, P., Power, C., & Ryan, A. M. Y. (2010). Teachers' education, teaching practice, and retention: a cross-genre review of recent research. *The Journal of Education, 191*(2), 19–31.

Cochran-Smith, M., Villegas, A. M., Abrams, L., Chavez-Moreno, L., Mills, T., & Stern, R. (2015). Critiquing teacher preparation research: an overview of the field, part II. *Journal of Teacher Education, 66*(2), 109–121. https://doi.org/10.1177/0022487114558268.

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Creemers, B., & Kyriakides, L. (2010). School factors explaining achievement on cognitive and affective outcomes: establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research, 54*(3), 263–294. https://doi.org/10.1080/00313831003764529.

Croninger, R. G., King Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*, 312–324.

Cunningham, A. E., & Zibulsky, J. (2009). Introduction to the special issue about perspectives on teachers' disciplinary knowledge of reading processes, development, and pedagogy. *Reading and Writing: An Interdisciplinary Journal, 22*, 375–378. https://doi.org/10.1007/s11145-009-9161-2.

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education, 51*(3), 166–173. https://doi.org/10.1177/0022487100051003002.

Darling-Hammond, L. (2014). Strengthening teacher preparation: the holy grail of teacher education. *Peabody Journal of Education, 89*, 547–561. https://doi.org/10.1080/0161956X.2014.93900.

Darling-Hammond, L. (2017). Teacher education around the world: what can we learn from international practice? *European Journal of Teacher Education, 40*(3), 291–309. https://doi.org/10.1080/02619768.2017.1315399.

Darling-Hammond, L., & Bransford, J. (Eds.). (2005). *Preparing teachers for a changing world. What teachers should learn and be able to do*. San Francisco: Jossey- Bass.

Darling-Hammond, L., Holtzman, d. J., Jin Gatlin, S., & Vasquez Heilig, J. (2005). Does teacher preparation matter? Evidence about teacher certification, teach for America, and teacher effectiveness. *Education Policy Analysis, 13*(42), 1–60. https://doi.org/10.14507/epaa.v13n42.2005.

Day, C. (2007). School reform and transitions in teacher professionalism and identity. In T. Townsend & R. Bates (Eds.), *Handbook of teacher education. Globalization, standards and professionalism in times of change* (pp. 597–612). The Netherlands: Springer.

Depaepe, F., Verschaffel, L., & Keltermans, G. (2013). Pedagogical content knowledge: a systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education, 34*, 12–25. https://doi.org/10.1016/j.tate.2013.03.001.

Dovemark, M., & Holm, A-S. (2017). The performative culture in Swedish schools and how teachers cope with it. In Borgnakke, K., Dovemark M., & da Silva, S.M. (Eds.), The postmodern professional. Contemporary learning practices, dilemmas and perspectives (pp. 33–52). London: Tufnell Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum.

Evetts, J. (2006). Short note: the sociology of professional groups. New directions. *Current Sociology, 54*(1), 133–143. https://doi.org/10.1177/0011392106057161.

Evetts, J. (2011). A new professionalism? Challenges and opportunities. *Current Sociology, 59*(4), 406–422. https://doi.org/10.1177/0011392111402585.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. https://doi.org/10.1080/00273170701382864.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001.

Ferguson, R. F. (1991). Paying for public education: new evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465–498.

Foy, P. (2013). *TIMSS and PIRLS 2011 User Guide for the Fourth Grade Combined International Database*. Chestnut Hill: Boston College.

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability, 30*, 159–182. https://doi.org/10.1007/s11092-018-9277-5.

Goe, L. (2007). *The link between teacher quality and student outcomes: a research synthesis*. Washington DC: National Comprehensive Center for Teacher Quality.

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: golden rules and editorial policies. *Personality and Individual Differences, 42*, 831–839. https://doi.org/10.1016/j.paid.2006.09.019.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *Review of Economics and Statistics, 89*, 134–150. https://doi.org/10.1162/rest.89.1.134.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement. An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*, 5–28. https://doi.org/10.1007/s11092-013-9179-5.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466–478. https://doi.org/10.1016/j.econedurev.2010.12.006.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *The Annual Review of Economics, 4*, 131–157. https://doi.org/10.1146/annurev-economics-080511-111001.

Hattie, J. A. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hedges, L. V., & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 74–92). Washington, DC: Brookings.

Hill, P. W., & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7*(1), 1–34. https://doi.org/10.1080/0924345960070101.

Hox, J. (2002). *Multilevel analysis - techniques and applications*. New Jersey: Lawrence Erlbaum Associates.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Imig, D. G., & Imig, S. R. (2006). The teacher effectiveness movement: how 80 years of essentialist control have shaped the teacher education profession. *Journal of Teacher Education, 57*(2), 167–180. https://doi.org/10.1177/0022487105285672.

Ingersoll, R. M. (2005). The problem of underqualified teachers: a sociological perspective. *Sociology of Education, 78*(2), 17–178.

Ingersoll, R. M. (2007). *A comparative study of teacher preparation and qualifications in six nations*. Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Ingersoll, R., Merrill, E., & May, H. (2016). Do accountability policies push teachers out? *Educational Leadership, 73*, 44–49.

Johansson, S., Myrberg, E., & Rosén, M. (2015). Formal teacher competence and its effect on pupil reading achievement. *Scandinavian Journal of Educational Research*, *59*(5), 564–582. https://doi.org/10.1080/00313831.2014.965787.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: the role of structural differences in teacher education. *Journal of Teacher Education, 64*(1), 90–106. https://doi.org/10.1177/0022487112460398.

Kline, R. B. (2016). *Principles and practices of structural equation modeling* (fourth ed.). New York: The Guilford Press.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583.

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*, 12–23. https://doi.org/10.1016/j.tate.2008.06.001.

Kyriakides, L., Creemers, B. P. M., Panayiotou, A., Vanlaar, G., Pfeifer, M., Cankar, G., & McMahon, L. (2014). Using student ratings to measure quality of teaching in six European countries. *European Journal of Teacher Education, 37*(2), 125–143. https://doi.org/10.1080/02619768.2014.882311.

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(010308), 1–29 Teachers College, Columbia University.

Lilja, P. (2014). A quest for legitimacy: on the professionalization policies of Sweden's teachers' unions. *Journal of Education Policy, 29*(1), 86–104. https://doi.org/10.1080/02680939.2013.790080.

Lindqvist, P., Nordänger, U. K., & Landahl, J. (2009). Insurance and assurance: teachers' strategies in the regime of risk and audit. *European Educational Research Journal, 8*(4), 508–519. https://doi.org/10.2304/eerj.2009.8.4.508.

Loehlin, J. C. (2004). *Latent variable models. An introduction to factor, path and structural analysis* (fourth ed.). Mahwah: Lawrence Erlbaum Publishers.

Luyten, H. (2003). The size of school effects compared to teacher effects: an overview of the research literature. *School Effectiveness and School Improvement, 14*(1), 31–51. https://doi.org/10.1076/sesi.14.1.31.13865.

Luyten, H., Visscher, A., & Witziers, B. (2005). School effectiveness research: from a review of the criticism to recommendations for further development. *School Effectiveness and School Improvement, 16*(3), 249–279. https://doi.org/10.1080/09243450500114884.

Markland, D. (2007). The golden rule is that there are no golden rules: a commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences, 42*, 851–858. https://doi.org/10.1016/j.paid.2006.09.023.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling, 11*, 320–341. https://doi.org/10.1207/s15328007sem1103_2.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2003). *PIRLS 2001 Technical Report*. Chestnut Hill: Boston College.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.

Myrberg, E. (2007). The effect of formal teacher education on reading achievement of 3rd-grade students in public and independent schools in Sweden. *Educational Studies*, *33*(2), 145–162. https://doi.org/10.1080/03055690601068311

Myrberg, E., Johansson, S., & Rosén, M. (2018). The Relation between Teacher Specialization and Student Reading Achievement. Scandinavian Journal of Educational Research, 1–15. https://doi.org/10.1080/00313831.2018.1434826

Nortvedt, G. A., Gustafsson, J.-E., & Lehre, A.-C. W. (2016). The importance of instructional quality for the relation between achievement in reading and mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes* (pp. 97–113). IEA Research for Education (A Series of Indepth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA)), vol 2. Springer, Cham

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257. https://doi.org/10.3102/01623737026003237.

Panayiotou, A., Kyriakides, L., Creemers, B. P. M., McMahon, L., Vanlaar, G., Pfeifer, M., …, Bren, M. (2014). Teacher behavior and student outcomes: results of a European study. Educational Assessment, Evaluation and Accountability, 26(1), 73–93. doi:https://doi.org/10.1007/s11092-013-9182-x.

Perry, L. A. (2006). Risk, error and accountability: improving the practice of school leaders. *Educational Research for Policy and Practice, 5*(2), 149–164. https://doi.org/10.1007/s10671-006-9002-x.

Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker—school league tables and English and Mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies, 59*(2), 179–195. https://doi.org/10.1080/00071005.2011.578568.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: evidence from panel data. *American Economic Review, 94*, 247–252.

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: an application of multilevel bifactor structural equation modeling. *Frontiers in Psychology, 6*(1550), 1–15. https://doi.org/10.3389/fpsyg.2015.01550.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317.

Snow, C. E., Griffin, P., & Burns, M. S. (2005). *Knowledge to support the teaching of reading. Preparing teachers for a changing world*. San Francisco: Jossey-Bass, The National Academy of education.

Stenlås, N. (2009). *En kår i kläm – Läraryrket mellan professionella ideal och statliga reformideologier. [A profession caught in the middle- teachers between professional ideals and state reform ideologies] ESO rapport 2009:6*. Stockholm: Regeringskansliet, Finansdepartementet.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: a review. *Review of Educational Research, 73*(1), 89–122. https://doi.org/10.3102/00346543073001089.

Zeichner, K. (2010). Competition, economic rationalization, increased surveillance, and attacks on diversity: neo-liberalism and the transformation of teacher education in the U.S. *Teaching and Teacher Education, 26*(8), 1544–1552. https://doi.org/10.1016/j.tate.2010.06.004.



## Affiliations

**Stefan Johansson**[1] · **Eva Myrberg**[1]

✉ Stefan Johansson
Stefan.johansson@gu.se

[1] Department of Education and Special Education, University of Gothenburg, Box 300, 40530 Gothenburg, Sweden