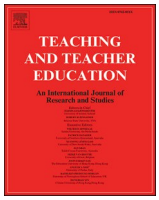
Teaching and Teacher Education 139 (2024) 104457

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate



Research paper

The impact of observable and perceived features of instruction on student achievement

Michael Tengberg [a,*], Gustaf B. Skar [b], Alan Huebner [c]

[a] Karlstad University, Sweden
[b] Norwegian University of Science and Technology, Norway
[c] University of Notre Dame, USA

ARTICLE INFO

Keywords:
Classroom observation
Language arts
Reading achievement
Student survey
Teaching quality

ABSTRACT

This study explored to what extent observable and perceived features of language arts instruction could explain variance in student reading achievement in lower secondary school. Data from classroom observations (using PLATO) and student surveys (using Tripod) were collected to examine the relationships between dimensions of teaching and student achievement gains (N=601). A combination of instructional features that provide coherence and consolidation of new knowledge were found to be positive predictors of reading achievement. The majority of features examined, however, did not significantly explain variance in achievement. We discuss the findings with respect to expected theoretical assumptions and potential measurement limitations.

1. Introduction

Teaching quality is critical to student learning (Burroughs et al., 2019; Charalambous et al., 2019; Hattie, 2009; Seidel & Shavelson, 2007). Over the last decade, a number of studies have suggested a positive relationship between teaching quality and student achievement (Blömeke & Olsen, 2019; Kane & Staiger, 2012; Kyriakides et al., 2013; Nilsen & Gustafsson, 2016; Praetorius et al., 2018). This effect is sometimes found to be stronger even than that of socioeconomic background, class size, or teachers' experience or training (Allen et al., 2011; Bryk et al., 2010; Hanushek, 2020; Konstantopoulos & Chung, 2011; Seidel & Shavelson, 2007). In order to explain classroom-level variance in achievement, both interventional research and observational studies of teaching have tried to identify key features of instruction, and to understand how they relate to one another (Allen et al., 2011; Creemers & Kyriakides, 2006; Muijs & Reynolds, 2018). Among the factors suggested by many scholars to be critical are teachers' classroom management, cognitive activation of tasks and activities, differentiation of instruction, supportive climate, and instructional clarity (Baumert et al., 2010; Hattie, 2009; Muijs et al., 2014; Praetorius et al., 2018; Senden et al., 2021). However, despite existing similarities in how quality of teaching is conceptualized and identified, there is still a great deal of uncertainty and debate about which features of practice are more important than others (Kyriakides et al., 2013; Senden et al., 2021), and to what extent positive relationships between specific teaching practices and student achievement extend across educational contexts (Blömeke & Olsen, 2019; Luoto, 2023), or even across different school settings (Cohen & Brown, 2016; Gill et al., 2016).

While many of the studies indicating a correlation between teaching and learning are based in an American educational context, e.g., the large Measures of Effective Teaching (MET) study (Kane & Staiger, 2012), there have been very few similar studies conducted in the Nordic countries. In the present study, we examine to what extent two US developed instruments (The Protocol for Language Arts Teaching Observation, PLATO, and The Tripod Student Survey) are able to explain variance in student achievement, using a sample of Swedish 7th grade language arts classrooms. PLATO (Grossman, 2015) is a language arts-specific observation system, whose components resonate well with prior research on Nordic language arts instruction and with language arts curricula in the region. Tripod (Ferguson, 2015) is a subject-generic instrument that was used as a complement to PLATO in the MET study (Kane & Staiger, 2012). As these two instruments have also been increasingly used to assess quality of teaching in Nordic lower secondary classrooms (Klette et al., 2021; Luoto et al., 2023; Tengberg et al., 2022), all the while there is still little evidence that the features captured actually predict achievement in Nordic classrooms, we include both of the instruments in the present study. Data for the study was collected from a sample of Swedish 7th grade (students aged 13–14 yrs) language arts classrooms, where student achievement gains were measured by

* Corresponding author. Department of Educational Studies, Karlstad University, 651 88 Karlstad, Sweden.
E-mail address: michael.tengberg@kau.se (M. Tengberg).

https://doi.org/10.1016/j.tate.2023.104457
Received 30 May 2023; Received in revised form 9 October 2023; Accepted 20 December 2023
Available online 3 January 2024

Abbreviations

PLATO The Protocol for Language Arts Teaching Observation

standardized tests of reading comprehension at two points with one year interval. Similar to the American context, in which PLATO and Tripod were developed, Swedish language arts instruction center on developing students' reading, writing, and speaking skills. Language arts in Sweden also incorporate instruction on Swedish language and literature. Recent examinations of language arts instruction in Swedish lower secondary indicate that practices targeted by PLATO and Tripod, such as clarity of instructional scaffolding, qualitative feedback on performance focus on conceptual understanding, and a supportive and cognitively challenging classroom discourse are equally critical to teaching quality in Sweden, as they are in the USA (Tengberg, 2022; Swedish Schools Inspectorate, 2012). However, there are no prior studies that systematically link those practices to student achievement gains.

1.1. Observing features of teaching quality

From the assumption that the observable presence of key features, in a small but representative sample of a teacher's practice, is indicative of student learning, various frameworks have been developed to examine in closer detail the relationship between instructional practices and student learning (Danielson, 2007; Grossman et al., 2013; Hamre et al., 2013; Praetorius et al., 2018). These frameworks operationalize and codify theories of the relationship between teaching and learning for large-scale systematic utility (Bell et al., 2012; Praetorius & Charalambous, 2018). Through scale-based specification of instructional features that are expected to be critical for learning, observation scores are compared across contexts, and used to assess differences and similarities between individual teachers, or between educational systems (Charalambous et al., 2019; Kane & Staiger, 2012; OECD, 2020). Global observation systems such as Classroom Assessment Scoring System (CLASS) (see e.g. Hamre et al., 2013), Three Basic Dimensions (TBD) (Praetorius et al., 2018), and PLATO (Grossman, 2015) are used for scoring teaching across supposedly representative selections of lessons, and assume that average scores across lessons relate positively to student achievement (White & Klette, 2023). Other systems work rather by the rationale of purposeful quantification, in which meaningful chunks of teaching are decided on and then analyzed by quantifiable measures, such as the rate of a certain type of teacher questions during classroom talk (Kelly et al., 2020) or the frequency of specific types of student utterances (Kosh et al., 2018).

PLATO 5.0 (Grossman, 2015) intends to capture four supposedly critical dimensions of language arts instruction (Instructional Scaffolding; Disciplinary Demand; Representations and Use of Content; and Classroom Environment) by analyzing specifically, on a four-point scale, twelve different indicator variables, i.e., twelve distinct features of teachers' instruction. These features include for instance teachers' provision of strategy instruction, connection of new content to students' prior knowledge, and teachers' management of lesson time and student behavior (see Table 1 four full display of the PLATO 5.0 variables). Theoretically, PLATO builds on both socioconstructivist and cognitive approaches to learning (Bell et al., 2019; Luoto et al., 2023). Several of the variables included favor a high level of student engagement and room for student thinking and contribution to academic discourse. Other variables emphasize that the teacher provide conceptual depth and clarity, and that tasks and activities enable an intellectual challenge for students. Overall, PLATO draws on research about critical features of high-quality language arts instruction in middle and/or lower secondary school (Grossman, 2015).

To produce metrics of instructional quality, scores on indicator variables are averaged across a number of lessons of observed teaching for each individual teacher. These average scores are then used either as direct predictors of achievement or averaged into dimensional (factor) scores or into an aggregated PLATO mean score. In prior studies, teachers' modeling and strategy instruction (Cohen, 2018; Grossman et al., 2013), and their provision of feedback to students (Klette et al., 2021), are examples of single indicators that were found to significantly contribute to student achievement in American and Norwegian classrooms respectively. Grossman et al. (2014), using an abridged version of PLATO called PLATO Prime, examined the impact on achievement by dimensions and by PLATO mean score. PLATO Prime included six of the twelve indicator variables presented in Table 1, and formed three dimensions with two variables in each dimension. Depending on which outcome measure was used, one or two of the dimensions (in both cases Classroom Environment) and the PLATO mean score were shown to be positively related to achievement. However, to the best of our knowledge, no prior study has examined the association between the four theoretically defined dimensions of PLATO 5.0 and student achievement, nor verified the four dimensions empirically.[1] For this reason, we first explore the factor structure of the twelve indicator variables in the present dataset, and then use the factors suggested by factor analysis to estimate the relationship between teaching and students' reading achievement gains. In this way, the identified dimensions of PLATO will serve as one of two different ways of operationalizing teaching quality in this study. The other operationalization of teaching quality is the Tripod survey, which is responded to by students.

Table 1
Original dimensions and indicator variables of PLATO 5.0 (Grossman, 2015).

Dimensions	Indicator variables
Instructional Scaffolding	Modeling and Use of Models (MOD) Strategy Use and Instruction (SUI) Feedback (FDBK) Accommodations for Language Learning (ALL)
Disciplinary Demand	Intellectual Challenge (IC) Classroom Discourse (CD) Text-Based Instruction (TBI)
Representations and Use of Content	Representation of Content (ROC) Connections to Prior Knowledge (CPK) Purpose (PUR)
Classroom Environment	Behavior Management (BM) Time Management (TM)

1.2. Student perceptions of teaching

Surveys of student perceptions of teaching are increasingly being introduced in teacher evaluation systems (Doherty & Jacobs, 2015). Both because classroom observations are costly and time consuming, and because students observe their teachers on a daily basis over much longer periods of time, researchers have argued that student responses to well-designed survey items may provide an effective and reliable alternative measure of teaching quality (Ferguson & Danielson, 2014; Kyriakides, 2005; van der Scheer et al., 2019). In a synthesis of teacher effectiveness research, Goe et al. (2008) point out that student ratings may indeed offer useful and valid information about teaching quality, making them a recommended alternative source of data, but that their validity is sometimes questioned because of possible biases. Such bias may for example be related to students' age or academic level, their expected or actual grades, or that scores reflect teachers' popularity

[1] There are prior studies which examine or use the factor structure of PLATO (Cor, 2011; Lazarev & Newman, 2014), but like Grossman et al. (2014) they use a different, and reduced, set of indicator variables, which limits their value for validation of the later, expanded version of PLATO which is used in the present study.

rather than their teaching quality (see also Fauth et al., 2014). As with other self-report instruments, however, validity of student ratings will depend on its design and validation procedures (Goe et al., 2008).

To what extent student perceptions can predict student achievement remains a matter of ongoing discussion (Maulana & Helms-Lorenz, 2016; van der Lans, 2018). Studies have shown for instance that student ratings of teaching were a better predictor of reading and mathematics achievement than principal ratings or teacher self-ratings (Wilkerson et al., 2000), and that third graders' ratings of teachers' classroom management predicted science achievements, while ratings of the teachers' cognitive activation and supportive climate did not (Fauth et al., 2014). Maulana and Helms-Lorenz (2016) showed that student perception was a better predictor of student academic engagement than observations of teaching, even though the instruments used in the study were theoretically aligned. Kyriakides (2005) found that student perceptions of the teacher-student relationship and cooperation correlated with student achievement in mathematics and Greek language.

As for the association between student perceptions and classroom observations, studies have suggested that the use of several observers observing several lessons increases the reliability of observation measures (e.g., Charalambous et al., 2017; Hill et al., 2012; Praetorius et al., 2014), which, in turn, increases the association between student perceptions and observations (van der Lans, 2018). This is expected since student surveys consider not only the practice in a single or a few lessons, but the practices employed over longer periods of schooling. As for the Tripod survey, Ferguson and Danielson (2014) examined its correlation with classroom observations by the Framework for Teaching (FfT) and found the strongest association (about 0.25) between factors in the two instruments that focused on teacher control/management of student behavior, whereas all other correlations between the two frameworks were around 0.15 or lower. To the best of our knowledge, there has been no empirical examination of the association between Tripod and PLATO. The present study offers none either, mainly because that would require a more thorough and theoretically informed discussion about the conceptual association between the two frameworks, a discussion for which there is not room in this paper. We do, however, incorporate both of the two measures as potential predictors of student achievement.

The design of the Tripod Survey rests on the theoretical assumption that effective delivery of instruction consist of a combination of sufficient content knowledge, pedagogical knowledge, and an ability to connect with and support students on a personal level (Ferguson & Danielson, 2014). It contains 38 items, articulated as statements about teachers teaching (see the entire scale in Appendix A). Together they aim to capture seven different dimensions (factors) of teaching (Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate) each of which is expected to be critical for student learning. In Table 2, each dimension is briefly explained and related to a typical teacher behavior for high scores on the given scale.

The survey asks students to consider to what extent the character of a given teacher's teaching, in the present study the language arts teacher, corresponds to each statement. The response format is a five-point Likert scale ranging from Never (1) to Always (5). Studies examining the underlying structure of Tripod have indicated high levels of covariance between latent factors, and suggested that the intended seven dimensions may not be supported by empirical data, and that alternative factor solutions for understanding the contribution of Tripod might be more productive (Kuhfeld, 2017; Schweig, 2014; Wallace et al., 2016). However, the results differ somewhat between studies. Using data from the MET study, Wallace et al. (2016) found that a bifactor model, with a general dimension including all items and a specific dimension consisting of the Control items, provided the best model fit. Kuhfeld (2017), also using MET data, later found that a two-dimensional model, with Control as one dimension and all other items as a combined "Support" dimension, provided the best fit.

Both Wallace et al. (2016) and Kuhfeld (2017) also examined the predictive validity of Tripod employing their respective factor solutions. Wallace et al. found that both the general factor and the Control factor significantly related to achievements in mathematics (with a coefficient of 0.25 in both cases), while Kuhfeld reported that only the Control factor significantly related to achievements by a coefficient of 0.18 and 0.14 in math and English respectively. Ferguson and Danielson (2014), also using MET data, investigated the predictive validity of a different structure (including Control and Challenge as separate factors and all others combined as a Support dimension) and found Control and Challenge to be significant predictors of student achievement (including both math and language arts) with regression coefficients of 0.24 for Control and 0.15 for Challenge. Similar results were reported in the MET papers ((Kane & Cantrell, 2010); Kane et al., 2013), although the predictive value of Tripod for language arts achievement was more limited than for mathematics. To conclude that Tripod is a robust predictor of language arts achievement would thus require additional evidence.

To sum up, the Control dimension of Tripod seems at least to be both psychometrically divisible from the rest of the scale and a significant predictor of student achievement. As for the other components of the instrument, results from prior research are not entirely consistent. In order to identify the underlying structure of Tripod in the Swedish dataset, we therefore begin by examining four of the previously suggested factor structures by confirmatory factor analysis (CFA). The one that provides the best model fit for our data would be the one most appropriate for examining relationship to achievement.

Different from Tripod, we had no previously empirically tested factor structure of PLATO, why we used exploratory factor analysis (EFA)[2] instead of CFA to explore the relation between indicator variables of PLATO. In this way, we tried to identify the most appropriate dimensional structure of the two instruments for predicting achievement.

Table 2
The seven dimensions of the Tripod Survey (Ferguson, 2015).

Label	Descriptor
Care	Teachers who *care* are emotionally supportive and interested in students.
Control	Effective *control* entails developing a respectful, cooperative classroom climate with on-task behaviour.
Clarify	Teachers who *clarify* explain things clearly, provide informative feedback, and clear up confusion in order to make lessons understandable.
Challenge	Teachers who *challenge* students press them to think rigorously and to persist when experiencing difficulty.
Captivate	Teachers who *captivate* make learning interesting and relevant.
Confer	Teachers who *confer* talk with students as well as welcome and respect student perspectives.
Consolidate	Teachers who *consolidate* summarize and integrate learning.

1.3. Research questions

The study explores to what extent classroom observations by PLATO, and student perceptions using the Tripod survey can explain variance of student reading achievement gains in lower secondary school. In accordance with the theoretical assumptions underlying the two instruments, we specifically examine whether

1. an increase in teacher PLATO scores (by dimensions) will have a significant positive effect on students' reading achievement gains, and whether
2. an increase in classroom Tripod scores (by dimensions) will have a significant positive effect on students' reading achievement gains.

2. Method

The study employs a research design in which video observations of

[2] More precisely, we used principal component analysis (PCA), see below.

language arts instruction, student perceptions of teaching gathered through survey responses, and student reading achievement data in the beginning of seventh grade (T1) and in the beginning of eighth grade (T2) were collected in order to analyze the relationship between features of instruction and student learning. The extent to which observable and perceived features of instruction explained posttest (T2) variance in student achievement gains was analyzed using multilevel modeling (MLM), which accounts for the fact that students are nested in schools and classrooms, and also allows control for the effects of student pretest results and gender.

2.1. Participants

The participants of the study were 601 seventh grade students (256 females and 345 males) distributed across 36 classrooms and 15 schools in the southern part of Sweden. Average number of students per classroom were 16.69 (*SD*=4.9). The sampling of classrooms was conducted to incorporate variation of teaching practices. Thus schools that were approached represented a variety with reference to geography, locality within community, immigrant proportion,[3] students' socioeconomic status, ownership form, grade average, and national test average. Two schools declined participation. The distribution of the remaining 15 schools across the mentioned factors are shown in Table 3. As seen, schools are set within different types of communities, immigration proportion and higher education proportion are varied but on average somewhat low compared to national averages, grade and national test averages are well distributed and well matched to national averages. The proportion of students in charter schools are low compared to the national average.

The teachers included in the sample varied in age (mainly between 30 and 50 yrs with a mean of 43.3 yrs), and were generally well qualified and experienced teachers of Swedish (mean of 15.0 years in service). In addition, the amount of professional development they had attended for the past five years ranged from none to several different courses or programs.

2.2. Data collection

Each teacher was observed during 3–4 lessons depending on length of their lessons. A normal lesson would last 50 min. In terms of 15-min segments, which is the case unit of analysis for observations in this study, the average was 12.7 segments per teacher. All lessons were captured on video using a set-up with two cameras (front and back) and two microphones (ceiling in mid of room and on teacher). As supplement to videos, we also collected photos of study material, white board annotations, classroom props, and student products. In line with previous research using PLATO (Grossman et al., 2013; Klette et al., 2017), these data were factored into the PLATO scores. Teachers were explicitly instructed not to make any changes of their teaching plans, but rather to teach the content and by the methods which they had already planned for. Besides single reports of anxiety and some extra preparation because there were cameras in the room, we received no indications of teachers making changes of their instructional plans.

2.3. Measures

The observed language arts lessons (N=134) were divided into 15-min segments (N=435) and scored by the twelve PLATO indicators. All raters were trained and certified in order to obtain a minimum of 80% reliability. During regular meetings, reliability was monitored by jointly scoring videos and deciding on critical issues and scoring rules. About 40% of all lesson segments were scored by two raters and disagreements settled through discussion. All scoring by PLATO is based on a four-point scale with qualitative criteria for each step and for each of the twelve elements (see abbreviated descriptions of the four levels of criteria for each indicator variable in Appendix B). Generally, low scores (1–2) represent that there is "almost no evidence" (1) or "limited evidence" (2) of the instructional feature in focus. Equally, high scores (3–4) represent that the observer sees "evidence with some weaknesses" (3) or "consistent strong evidence" (4) for the feature in focus (Grossman, 2015).

To measure student perceptions of teaching, the Tripod Survey was distributed to students (N=601) after the final video-recording in each class respectively. Overall reliability was alpha=0.954 suggesting a high general consistency across the scale.

Reading comprehension was measured using a standardized reading test (Norwegian national reading test for 13–14 year-olds) in the beginning of 7th grade (T1) and beginning of 8th grade (T2). The test comprises a selection of seven texts and 43 (T1) or 44 (T2) items (75% multiple choice and 25% short answer constructed response). The text selection contains six non-literary texts from a broad variation of content areas and one narrative text. Items are designed by and distributed across the three cognitive approaches defined by the PISA framework: *access and retrieve*; *integrate and interpret*; and *reflect and evaluate*.

2.4. Descriptive statistics and factor reduction

Tables 4 and 5 provide descriptive statistics of reading comprehension (pre- and posttest) on school, classroom and student level. As seen in Table 4, there was a significant increase of reading comprehension from pretest to posttest (p=0.001). On the student level, the effect size (Hedges' g) amounted to 0.32, after taking the correlation between pretest and posttest results (r=0.79) into account. Table 4 also indicates that the dispersion of reading comprehension was greater on student level (*SD*=10.0) than on classroom (*SD*=4.4) and school level (*SD*=3.7). Girls outperformed boys at pretest (Hedges' g=0.52) and posttest (Hedges' g=0.46). Within-gender increase from pretest to posttest (displayed in Table 5) corresponded to g=0.34 and g=0.31 for girls and boys respectively.

To reduce the twelve indicator variables of PLATO into an appropriate factor structure, we applied principal component analysis (PCA). Results are displayed in Table 6, suggesting that the twelve variables load on four different factors, in a structure that closely resemble the original structure advocated by the developers of PLATO (Grossman, 2015). Three of the twelve variables load on factors other than the ones proposed by the original model (see Table 1). From a theoretical perspective, however, this adjustment is sensible. MOD, SUI, ALL, and ROC (Factor 1) all capture a teacher's demonstration of strategies, concepts, knowledge and procedures for how students can approach educational content. FB, IC, and TBI (Factor 2) have in common that they capture means for developing the quality of students' work, and for increasing the intellectual rigor and subject-related demands on students' task completion. CD, CPK, and PUR (Factor 3) concern aspects of the teacher's pursuits to consolidate and build coherence of the content, while BM and TM (Factor 4) concern the teacher's classroom management.

Scores on each individual PLATO indicator variables, and descriptive statistics for single Tripod items across all students in the sample, are provided in Appendix C.

Table 7 displays the correlations between PLATO indicator variables. In Factor 1, MOD, SUI, ALL, and ROC displayed positive correlations in the range r=0.26 to r=0.54. In Factor 2, IC, TBI, and FB also displayed positive correlations, in a somewhat narrower range (r=0.32 to r=0.54). Factor 3 (CPK, PUR, and CD) included two rather weak correlations (between CD and CPK, and between CD and PUR, r=0.17, and r=0.16 respectively. The correlation between CPK and PUR was r=0.33. The

[3] Schools in Sweden do not register data on 'ethnicity' or 'race', but on the proportion of students with immigrant background. A student is defined to have immigration background when s/he is *either* born outside of Sweden *or* is born in Sweden but both parents are born abroad.

Table 3
Characteristics of sample schools.

School no	School locality[a]	Immigr. proport.[b]	Higher ed. proport.[c]	Public/Charter[d]	Merit Aver.[e]	NT Aver. LA[f]	NT Aver. MA[g]
1	Town	0.16	0.64	Public	216.5	14.3	10.6
2	Village	0.07	0.37	Public	198.1	10.8	10.1
3	Town	0.08	0.71	Public	225.3	13.9	10.8
4	Town	0.08	0.73	Public	229.6	13.6	11.7
5	Town	0.69	0.37	Public	187.8	12.5	9.2
6	Small town	0.14	0.46	Public	205.9	12.5	10.2
7	Village	0.29	0.31	Public	217.2	13.1	12.4
8	Town	0.14	0.67	Public	242.1	15.3	12.4
9	Town	0.2	0.53	Public	201.5	13.4	10.3
10	Town	0.19	0.54	Public	201.0	12.4	9.0
11	Small town	0.37	0.25	Public	182.4	12.3	9.7
12	City	0.1	0.87	Charter	271.6	16.0	14.6
13	Town	0.14	0.66	Public	231.9	13.6	–
14	Town	0.23	0.51	Public	211.4	14.0	11.4
15	Town	0.29	0.46	Public	203.0	12.7	11.0
Sample average		**0.21**	**0.54**	**0.11**	**215.0**	**13.4**	**10.9**
National average		0.25	0.59	0.20	216.9	13.6	11.2

[a] In accordance with definitions in the PISA School Questionnaire (Item SC001) (OECD, 2017). [b] Proportion of students born abroad or both parents born abroad, data from the year of our data collection. [c] Proportion of students with at least one parent with tertiary education in the year of data collection. [d] Public or charter school. Average represents proportion of students in school year 7 in sample vs. in national population attending charter schools during the year of data collection. [e] Based on student grades in all subjects in 9th grade. Average of the 5 years preceding the data collection. [f, g] National test averages in 9th grade language arts and mathematics respectively. Scale from 0 to 20. Average of the 5 years preceding the data collection.

Table 4
Reading comprehension results for schools, classrooms, and students.

	N	Pretest Mean (SD)	Posttest Mean (SD)	p-value	Correlation	Effect Size
Schools	15	46.0 (3.7)	49.5 (4.4)	<.001	.85	0.84***
Classrooms	36	46.3 (4.4)	49.6 (5.2)	<.001	.83	0.68***
Students	601	46.8 (10.0)	50.2 (11.0)	<.001	.79	0.32***

Note. Effect size is Hedges' *g*.

Table 5
Reading comprehension results for girls and boys.

	N	Pretest Mean (SD)	Posttest Mean (SD)	p-value	Correlation	Effect Size
Girls	256	49.7 (9.2)	53.0 (10.3)	<.001	.79	0.34***
Boys	345	44.7 (10.1)	48.0 (11.0)	<.001	.76	0.31***
Effect Size		Pretest	Posttest			
Girls vs Boys		0.51***	0.47***			

Note. Effect size is Hedges' *g*.

correlation between the two items of Factor 4 (BM and TM) was r=0.44.

The present study focus on the impact on achievement by factors, or dimensions, of teaching rather than by single indicator variables. However, as several previous studies using PLATO examine impact by single variables (Cohen, 2018; Grossman et al., 2013; Klette et al., 2021; Lazarev & Newman, 2014), we provide, for comparative purposes, a display of the relative contribution to variation in the dependent variable by single indicator variables in Appendix D.

To verify the underlying structure driving student responses to Tripod and model the responses to the dimensions proposed by previous research, we conducted a series of confirmatory factor analyses (CFA). If the underlying factors are to be used as predictors of student achievement, the reasonable approach would be to identify the structure that best fits the present response data. Four alternative models, referred to in Table 8 as MOD1–4, were examined based on prior research. It should

Table 6
Factor loadings of indicator variables.

	Factor 1	Factor 2	Factor 3	Factor 4
Modeling and Use of Models (MOD)	.808			
Strategy Use and Instruction (SUI)	.740			
Feedback (FDBK)		.657		
Accommodations for Language Learning (ALL)	.712			
Intellectual Challenge (IC)		.837		
Classroom Discourse (CD)			.656	
Text-Based Instruction (TBI)		.785		
Representation of Content (ROC)	.606			
Connections to Prior Knowledge (CPK)			.684	
Purpose (PUR)			.669	
Behavior Management (BM)				.842
Time Management (TM)				.831

be noted that all four models derive from different groupings of the originally proposed seven factors by Ferguson (2015).

The R package lavaan (Rosseel, 2012) was used to perform the analyses. The statistical criteria Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and a 95% confidence interval (CI) for the RMSEA are displayed in Table 9. The CFI and TLI should be maximized, where the rules of thumb indicating good fit are >0.95 and >0.90 for the CFI and TLI, respectively. The RMSEA should be minimized, where the thresholds are <0.05 indicates excellent fit and between 0.06 and 0.09 indicates adequate fit. According to CFA, MOD1, the seven factor model, shows the best fit of all four models: the lower bound of the 95% CI for the RMSEA is just at the threshold of excellent fit. MOD4, the bifactor model, showed the best fit of the three other models, but the output indicated convergence issues, which could indicate that the model is incorrectly specified. Thus, MOD1, the seven factor model, was deemed as the most appropriate for studying prediction of student achievement.

Estimated correlations between the seven dimensions of Tripod suggested that these were highly associated with one another, something which is expected as the seven factors tap into related constructs (see Table 10). Care and Clarify are the ones most strongly correlated (0.920) with one another, while the Control and Captivate are the least correlated (0.419).

To sum up, the two factor analyses identify appropriate dimensional

Table 7
Correlations between PLATO variables.

	MOD	SUI	ALL	ROC	IC	TBI	FB	CPK	PUR	CD	BM
MOD	–										
SUI	.54	–									
ALL	.33	.30	–								
ROC	.44	.31	.26	–							
IC	−.04	.05	.05	−.05	–						
TBI	−.09	−.02	−.01	−.11	.54	–					
FB	.06	.12	.10	.25	.36	.32	–				
CPK	.14	.23	.10	.23	−.20	−.10	.00	–			
PUR	.18	.16	.07	.27	−.06	−.12	.04	.33	–		
CD	.11	.13	−.03	.17	.08	−.05	.09	.17	.16	–	
BM	−.01	−.04	−.05	.11	−.05	−.08	.05	−.02	.09	.05	–
TM	−.01	−.01	.00	.04	.08	.03	.08	−.06	.00	.04	.41

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Rotation converged in 5 iterations.

Table 8
Alternative factor structures estimated by CFA.

Model	Description and reference
MOD1	Seven factors (the original structure proposed by Ferguson, 2015 and displayed in Table 2)
MOD2	Two factors: Press (including Control and Challenge) and Support (including Care, Clarify, Captivate, Confer, and Consolidate) (proposed by Ferguson & Danielson, 2014)
MOD3	Two factors: Control and Composite (including Care, Clarify, Challenge, Captivate, Confer, and Consolidate) (identified by Kuhfeld, 2017)
MOD4	Bifactor model consisting of a General factor and Control as a separate factor (suggested by Wallace et al., 2016)

Table 9
Fit criteria for MOD1-4. The 95% CI for the RMSEA is given in parentheses.

MOD	CFI	TLI	RMSEA
MOD1	0.914	0.906	0.054 (0.050, 0.057)
MOD2	0.782	0.770	0.084 (0.081, 0.087)
MOD3	0.852	0.843	0.070 (0.066, 0.073)
MOD4	0.856	0.846	0.069 (0.066, 0.072)

Table 10
Correlations among seven latent factors.

	Care	Control	Clarify	Challenge	Captivate	Confer
Control	.506					
Clarify	.920	.492				
Challenge	.844	.449	.873			
Captivate	.804	.419	.778	.754		
Confer	.917	.466	.854	.851	.815	
Consolidate	.859	.433	.856	.883	.722	.861

structures of the two instruments for describing related aspects of teaching, and for explaining variance in achievement. The four identified dimensions of PLATO, and the seven confirmed dimensions of Tripod will thus be used as predictors in the subsequent analysis.

2.5. Analytical strategy

The relationships between student demographics, teaching variables (PLATO and Tripod) and reading comprehension scores were examined using multilevel models (MLMs), which account for the multiple levels of clustering. Pretest and posttest scores of reading are nested within students, which are in turn nested within classrooms and schools, resulting in a four-level MLM. Thus, the MLM framework allows for the examination of the correlation induced by the clustering as well as ensures valid statistical inferences. A number of MLMs were fit, and for each model, the response variable is reading score.[4]

- Null Model: First, the null model (i.e., empty model with no predictors) was fit with reading comprehension score as the dependent variable, and the intraclass correlations (ICCs) were computed to assess the correlation due to the clustering.
- Model 1: The means of the four PLATO factors were entered as independent variables, and gender (1=male, 0=female) and time (1=posttest, 0=pretest) were entered as binary independent variables. Thus, in total six covariates were considered.
- Model 2: The means of the seven Tripod factor scores, along with gender and time, were used as predictors.

For each of Models 1, 2, and 3, single covariates or factors were removed and the model was refit, and the difference in R^2 between the full model and the smaller model was interpreted as the contribution of the covariate or factor that was removed.

3. Results

3.1. Null model

The ICCs were computed by extending the method of Snijders and Bosker (2012, p. 68−69) for MLMs with three levels, and the values are reported in Table 11. Various ways to interpret an ICC are reviewed by Lorah (2018) and Snijders and Bosker (2012). The student-level ICC is 0.74, indicating that the estimated correlation of pretest and posttest reading scores for a randomly selected student is 0.74. Furthermore, the estimated correlation for scores obtained from two students randomly selected from the same classroom is 0.13, and the estimated correlation for scores obtained from two randomly selected students in the same school is 0.08. Thus, as expected, the relation between students in classrooms is slightly stronger than the relation between students in

Table 11
Intraclass correlations.

Level	Variance Component	Corresponding ICC
Time (Level 1)	29.77	–
Student (Level 2)	70.49	.74
Classroom (Level 3)	5.60	.13
School (Level 4)	9.25	.08

[4] All MLM analyses were performed using the lme4 package (Bates et al., 2015) in R.

schools.

3.2. Model 1

As described above, Model 1 was fit with 6 covariates; the model including all of them will subsequently be referred to as the "full model". The four continuous PLATO factors were scaled to have a mean of zero and standard deviation of one to make their magnitudes comparable amongst themselves and across other studies with similar populations (Lorah, 2018). The regression effect estimates are presented in Table 12. Looking at the effect estimates for Time and Gender, respectively, we expect the posttest mean score to be about 3.341 points higher than the pretest, and girls' mean score to be about 4.847 higher than the boys' score, on average. These effects are highly significant ($p<0.001$). Among the PLATO factors, Factor 3 (including Connections to Prior Knowledge (CPK), Purpose (PUR), and Classroom Discourse (CD)) is highly significant ($p=0.002$). The slope estimate of 2.944 suggests that, for every standard deviation increase in the factor, we expect the reading comprehension score to increase by almost 3 points, on average.

While the ICC provides information about the correlation structure of the data arising from the clustering, it does not tell us anything about the explanatory, or predictive, power of the covariates. Thus, R^2 is a measure needed to indicate the amount of variation in the dependent variable that is explained by the covariates. The R^2 for this full model is 0.126, while the amount of R^2 contributed by the four factors is reported in Table 13. For example, the first line reports that a new model was fit without the covariate Time, and the resulting R^2 was 0.102, meaning that Time explains approximately 2.4% of the variation in reading score. Then, a model was fit without Gender (Time was put back into the model) and the R^2 of that model was 0.079, and so on. Of the PLATO factors, we see that Factor 3 contributes the most R^2 (i.e., Factor 3 itself explains nearly 4.5% of the variation in the outcome). This is consistent with the fact that Factor 3 was the only significant factor in the regression reported in Table 12. We see that Factor 4 contributes negatively to R^2, i.e., the explained variation decreases when the elements BM and TM are added (or increases when they are removed). This decrease in R^2 is, however, likely due to random fluctuation (cf., Snijders & Bosker, 2012, p. 112–113, 156).

3.3. Model 2

Model 2 included as covariates the means of the seven Tripod factor scores, along with time and gender. Similar to the continuous covariates in the previous models, the Tripod scores were put on a standard scale with mean of 0 and standard deviation of 1. The regression results are shown in Table 14. Again, time and gender have highly significant ($p<0.001$) regression effects. However, among the Tripod factors, we identify no significant effects. Several of the reported effects are negative, suggesting a negative influence of these teacher practices should they have been significant. But the generally high p-values indicate that the observed effects are only due to random sampling variability.

The amount of R^2 contributed by the seven Tripod factors is reported in Table 15. As shown, these contributions (to the full model at 0.082) are all minor but positive. The strongest contribution of the seven factors is provided by Clarify, amounting to 2% of the variance in students' posttest reading scores.

Table 12
Regression results for the four-factor model for the PLATO data.

	Estimate	95% CI	P-value
Intercept	49.307	(47.389, 51.225)	<.001
Time	3.341	(2.783, 3.89)	<.001
Gender	−4.847	(-6.321, −3.373)	<.001
Factor 1	−0.064	(-1.822, 1.695)	.941
Factor 2	−0.370	(-2.067, 1.327)	.659
Factor 3	2.944	(1.199, 4.689)	.002
Factor 4	−0.211	(-1.727, 1.306)	.778

Note. Continuous covariates are standardized to have mean of 0 and standard deviation of 1.

Table 13
Contributed R^2 for Time, Gender, and PLATO factors.

Variable(s) Removed	R^2	Difference from full model (i.e. R^2 contributed by Variables)
Time	.102	.024
Gender	.079	.047
Factor 1	.108	.018
Factor 2	.113	.013
Factor 3	.083	.043
Factor 4	.136	−.010

Table 14
Regression results for Model 2.

	Estimate	95% CI	P-value
Intercept	49.219	(46.954, 51.484)	<.001
Time (Posttest)	3.341	(2.783, 3.899)	<.001
Gender (Boy)	−4.888	(-6.363, −3.413)	<.001
Care	−0.749	(-5.603, 4.104)	.753
Control	−0.491	(-2.651, 1.669)	.642
Clarify	3.502	(-0.777, 7.781)	.104
Challenge	−0.057	(-3.254, 3.141)	.971
Captivate	−1.430	(-5.187, 2.328)	.439
Confer	2.173	(-1.295, 5.640)	.209
Consolidate	−2.096	(-6.751, 2.560)	.363

Note. Continuous covariates are standardized to have mean of 0 and standard deviation of 1.

4. Discussion and implications

In response to a growing interest in identifying factors of teaching quality, it is critical to examine the validity, including the cross-contextual validity, of measures expected to capture the relationship between teachers' performance and student learning. Attempts over the past decades to conceptualize and operationalize teaching quality has resulted in a vast number of frameworks aiming to identify the most relevant factors for teaching in general or for teaching a specific subject (Bell et al., 2019; Praetorius & Charalambous, 2018; Seidel & Shavelson, 2007). Because estimates of teachers' different contribution to learning vary between frameworks and contexts (Blömeke & Olsen, 2019; Cohen & Brown, 2016; Senden et al., 2021), additional research is clearly needed to gauge the leverage of various factors within different educational contexts and in relation to different outcomes. The present study investigated to what extent classroom observations by PLATO, and student perceptions using the Tripod survey can explain variance of student reading achievement in Swedish lower secondary school. Specifically, we examined whether 1) an increased PLATO score (by

Table 15
Contributed R^2 for Time, Gender, and Tripod factors. The R^2 for Model 3 is 0.082.

Variable Removed	R^2	Difference from full model (i.e. R^2 contributed by Variables)
Time	.046	.036
Gender	.019	.063
Care	.076	.006
Control	.072	.010
Clarify	.064	.018
Challenge	.074	.008
Captivate	.074	.008
Confer	.070	.012
Consolidate	.072	.010

dimensions) had a significant positive effect on students' reading achievement gains, and whether 2) an increased Tripod score (by dimensions) had a significant positive effect on students' reading achievement gains.

Based on observation data, the combination of three instructional features were found to impact the development of reading comprehension significantly: Connections to Prior Knowledge (CPK), Purpose (PUR), and Classroom Discourse (CD). A teacher scoring high on CPK connects new material to students' previous academic knowledge by referring to prior lessons or eliciting student knowledge in class. The importance of such linkage to student learning has been accounted for in several prior studies (Scott et al., 2011; Silseth & Erstad, 2018). Scoring high on PUR means to provide students with situated and internal learning goals for the classroom activity, i.e., to provide both context and specifications for what students are expected to learn. From prior studies, we know that Swedish teachers are generally explicit about what students are expected to *do* during the lesson, but not very explicit about what they are expected to *learn* from the activities (Tengberg, 2022). The significance of engaging students by making future relevance specific, and by facilitating learning through addressing learning goals explicitly has been verified in other studies (Ames, 1992; Locke & Latham, 2002; Pintrich, 2000; Spinath & Steinmayr, 2012). A high score on CD means to provide ample opportunities and strategic support for students' content-related talk in the classroom. CD highlights authentic questioning, uptake of student responses, and extended room for students to verbalize their understanding of content, qualities that prior research also associates with increased reading comprehension (Murphy et al., 2009; Wilkinson et al., 2015). Based on evidence from prior research, it is reasonable to infer that the combination of CPK, PUR, and CD provides coherence and consolidation of new knowledge, thereby giving students opportunity to put learning into context and envision the cognitive objectives of their efforts. According to the results, the combination of these three instructional features (Factor 3) explained 4.3% of the posttest variance, which is almost on par with the effect of gender. Since the gender effect (girls advantage over boys) in this sample was equal to g=0.47 on the posttest, the joint impact of CPK, PUR, and CD is substantial enough to be of clear pedagogical interest, and worthy of further study. In addition, the Tripod factor Clarify, targeting the clarity of teachers' explanations of content, and their ability to adjust pace and level of instruction to students' understanding, may explain an additional 2% of the posttest variation, although the estimated effect was not statistically significant (p=0.104).

None of the other PLATO or Tripod factors, however, significantly impacted students' reading achievement. Prior studies using PLATO to predict achievement have mainly examined the effects of single indicator variables (Cohen, 2018; Cohen & Grossman, 2016; Klette et al., 2021), and we have found no study that uses factor scores from the full PLATO model (PLATO 5.0) to explain variance in reading achievement. Lazarev and Newman (2014) defined three factors from eight of the single PLATO variables and investigated various approaches of correlating observation scores with student outcomes. The contributions of factor scores to student gains in their model was represented by regression coefficients in the range of 0.04–0.05. One of the factors, including the variables Behavior Management (BM) and Time Management (TM), coincides with Factor 4 in our study. While our results were non-significant, Lazarev & Newman found a significant and positive contribution of this factor. Their finding also align with prior research indicating that students' perception of teacher control and classroom management associate positively with teacher value-added scores (Kuhfeld, 2017; Wallace et al., 2016).

One interpretation of the results is that teaching quality is not a phenomena that easily translates into a generalized index of many different variables, presumably because teachers are skilled at different things and few teachers outclass their colleagues at a majority of the variables included. In a previous study, we found that while there was large variation between classrooms on single observation variables, average PLATO scores were much less varied (Tengberg et al., 2022). We also found (although non-significant) indications of a differential impact of PLATO factors on boys and girls respectively, which may contribute to explain a lack of significant effect on the whole sample. Although this study was able to identify factors related to variance in student outcome, a substantial proportion of variance still remains unexplained. If we are to maintain the belief that variation in teaching quality contribute to variation in student learning over time, we need to address this lack of explanatory power. In the following, we identify some of the limitations in the present study in order both to emphasize caution about extrapolating from the study, and to highlight possible ways forward in the scientific study of the relationship between teaching quality and student learning.

4.1. Sample size

Compared with samples included in similar prior studies (Allen et al., 2011; Cohen, 2018; Grossman et al., 2014; Klette et al., 2021), the sample size in the present study is relatively small (36 classrooms). A smaller sample is more vulnerable to random error of the measurement, and limits the likelihood of locating statistically significant relationships. However, since most of the non-significant relationships found in the study were also weak, it is not certain that a larger sample size, all else being equal, would have changed the overall picture of the association between observed teaching and student learning. What a larger sample might do, on the other hand, is to yield larger variation on the variables examined, which might in turn provide better opportunity to model statistical relationships. However, to increase sample size is both costly and time consuming. In a Nordic context, an observation study of teaching quality based on 134 lessons in 36 different classrooms, including measures of student perception and gains, is comparatively large. Few studies investigating the quality of teaching in Nordic schools reach this size. To collect larger samples also renders a number of practical challenges that has to do with research infrastructures and the decentralized structure, and extensive autonomy, of schools and teachers. Yet, the need for a larger sample also relates to the ambition of generalizing the concept of teaching quality across methods, purposes, contexts (including students), and content. We address some additional concerns that relate to this problem below.

4.2. Outcome measure

In the study, measurement of reading comprehension was used for estimating student achievement across one school year. Although there is evidence that reading predicts achievement in several other areas (Childs et al., 2014; Ritchie & Bates, 2013), it is uncertain whether all of the features captured by PLATO contribute specifically to students' reading comprehension. PLATO is defined by features expected to be critical in language arts teaching, and several of the variables (including Feedback, Intellectual Challenge and Connections to Prior Knowledge) may certainly be regarded as even more generic still, suggesting that PLATO is not even a completely subject-specific but a *hybrid* framework (Charalambous & Praetorius, 2018), i.e., attending to both general and subject-specific aspects of teaching. Moreover, teaching of language arts promotes not only reading but a range of different skills including literature, writing, oracy, and language. In addition to broad measures of teaching, linking teaching to achievement might therefore require broader measures of student achievement to provide better estimates. This point was referred to by Grossman et al. (2013) when emphasizing that the features of instruction captured by PLATO may very well be important for the development of many other language arts-specific skills than the ones tested for constructing value-added scores. However, more extensive packages of testing may be both resource-demanding and contra-productive to teachers' willingness to participate in research. An alternative research design strategy could be to form closer connections between the type of teaching observed in the classroom, or

asked about in surveys, and the specific measures of student learning.

A related aspect concerns the repeated measurement of student achievement. Ideally, an estimation of students' reading growth, as an effect of teaching, would include several time points for measuring achievement. In the present study, we were unable to measure reading comprehension more than two times. In addition, the present study did not include controls for student characteristics, as was the case for example in the MET study (Kane & Staiger, 2012). To some extent, pre-test scores reflect student background, but since background factors may also influence the rate at which students learn (Ballou et al., 2004), such data would have improved the study's robustness. Future investigations of teachers' impact on learning should thus seek to include measures of achievement across several time points in order to estimate growth curves. They should also try to include measures of student characteristics (e.g., socio-economic status, language background, and achievement in other subjects) in order to exert more precise control in the estimates of how teaching practices affect changes in reading comprehension.

4.3. Content dependence of observation scores

A prior analysis of the present sample showed that some of the PLATO variable scores were related to content, for example that teachers gave significantly more feedback (FB) during writing instruction than during instruction in other content areas (Tengberg, 2022). Grossman et al. (2010) also reported significantly different scores in different content areas. Since our data, for practical reasons, were collected during four consecutive lessons rather than during lessons dispersed across the school year, the teaching observed in the different classrooms did not include an even distribution of content between classrooms. Therefore it is hard to entirely disentangle the influence of content from the quality of teachers' performance. If, for instance, qualitative feedback is more common during writing instruction than during reading instruction, we cannot infer that the individual teachers who happened to devote more time to writing than reading during our observations generally provided more qualitative feedback to their students. This may therefore be a source of measurement error, also indicating the need for a larger and more strategically collected sample.

4.4. Theoretical assumption of linear relationship

Finally, it is worth reflecting over the assumption of linear relationship between observation scores (or student perceptions) and student gains. Lazarev and Newman (2014) found that several of the single PLATO variables had non-linear (both U-shaped and S-shaped) relationships with the student outcomes, suggesting that ideal averages for those variables were not necessarily at the top of the distribution. The present study contained no plotting of the functional relationship between independent and dependent variables, but it is necessary to consider whether a higher average score on PLATO or Tripod variables would always be expected to yield better opportunities for student learning (cf. White & Klette, 2023). While some features of instruction are clearly constantly critical, such as using lesson time effectively (TM) and avoiding disruptions that distract students from learning (BM), other features may be more important if delivered qualitatively and aptly at the right time rather than high on average. Operationalization of target features of instruction must therefore take into account whether a high score should represent the quality, the timing, or the frequency of applying a particular teaching strategy. As noted by Praetorius and Charalambous (2018), several observation frameworks intertwine quality and frequency when rating specific features, and while both quality and frequency may be of importance, they are not always aligned, which may entail measurement error. More generally, this concerns the possibility that some variables or dimensions of PLATO do not capture well enough the features of instruction they intend to capture (cf. Luoto et al., 2023). This should be considered carefully both in the interpretation of PLATO scores and in the design of new observation instruments.

5. Conclusion

Despite these challenges, the present study provides important insights about the ability to explain variance in student achievement by observations and student perceptions of instruction. PLATO was designed for American classrooms, and most of the research evidence about PLATO is produced by American studies. Since indices from cross-country comparisons point to non-trivial culture differences in teaching (OECD, 2018, 2020; Stigler & Hiebert, 1999; Xu & Clarke, 2018), it appears necessary to validate PLATO against relevant outcome measures in dissimilar educational contexts, in this case the Swedish, before assuming that findings produced in America are generally transferable. While the present study was able to identify a combination of instructional features (providing coherence and consolidation of new knowledge for students) that significantly contributed to explaining variance in student achievement, the study also suggests that associating more general dimensions of teaching quality with student learning over time seem to require more refined measurements than the ones applied here. Such studies would for example benefit from both larger samples and a narrower focus on types of teaching more directly related to specific student outcomes. Although prior large-scale research have suggested strong positive impact of teaching quality on student learning (Allen et al., 2011; Hanushek, 2020; Seidel & Shavelson, 2007), the present study was unable to verify any general or strong relationships using observations by PLATO and student perceptions by Tripod to operationalize teaching quality. Whether this result is attributable to factors of the Swedish educational context lies beyond the scope of the study. It seems clear, however, that additional evidence from larger samples is needed to provide stronger argument for the validity of PLATO and Tripod as operationalizations of teaching quality. The findings from the study may also point to a more general implication, namely that policy makers, school leaders, and teachers should be careful about expecting to find clear-cut or generalizable, linear relationships between observable features of teaching and the extent of student learning. This does not undermine the idea that the quality of teaching is critical to achievement, but it suggests that the relationship is more complicated than in the theoretical model proposed and examined in this study.

Declaration of competing interest

All authors declare that they have no relevant interests to disclose.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by The Swedish Research Council [grant number 2017-03544].

Appendices A-D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2023.104457.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancingsecondary school instruction and student achievement. *Science, 333*(6045), 1034–1037. https://doi.org/10.1126/science.1207998

Ames, C. (1992). Achievement goals and the classroom motivational climate. In D. H. Schunk, & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 327–348). Erlbaum.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87. https://doi.org/10.1080/10627197.2012.715014

Blömeke, S., & Olsen, R. V. (2019). Consistency of results regarding teacher effects across subjects, school levels, outcomes and countries. *Teaching and Teacher Education, 77*, 170–182. https://doi.org/10.1016/j.tate.2018.09.018

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organising schools for improvement: Lessons from Chicago*. University of Chicago Press.

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, & W. Schmidt (Eds.), *Teaching for excellence and equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). Springer International Publishing.

Charalambous, C. Y., Kyriakides, E., Kyriakides, L., & Tsangaridou, N. (2019). Are teachers consistently effective across subject matters? Revisiting the issue of differential teacher effectiveness. *School Effectiveness and School Improvement, 30*(4), 353–379. https://doi.org/10.1080/09243453.2019.1618877

Charalambous, C., Kyriakides, L., Tsangaridou, N., & Kyriakides, L. (2017). Exploring the reliability of generic and content-specific instructional aspects in physical education lessons. *School Improvement and School Effectiveness, 28*(4), 555–577. https://doi.org/10.1080/09243453.2017.1311929

Charalambous, C. Y., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: Setting the ground for understanding instructional quality more comprehensively. *ZDM, 50*(3), 355–366. https://doi.org/10.1007/s11858-018-0914-8

Childs, S., Finnie, R., & Mueller, R. E. (2014). The cultural determinants of access to post-secondary education by first generation households: An analysis using the youth in transition survey. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2463723.

Cohen, J. (2018). Practices that cross disciplines? Revisiting explicit instruction in elementary mathematics and English language arts. *Teaching and Teacher Education, 69*, 324–335.

Cohen, J., & Brown, M. (2016). Teaching quality across school settings. *The New Educator, 12*(2), 191–218. https://doi.org/10.1080/1547688X.2016.1156459

Cohen, J., & Grossman, P. (2016). Respecting complexity in measures of teaching: Keeping students and schools in focus. *Teaching and Teacher Education, 55*, 308–317.

Cor, K. (2011). *Investigating the reliability of classroom observation protocols: The case of PLATO*. New Orleans, LA: Paper presenterat vid The Annual Meeting of the American Educational Research Association.

Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modeling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347–366.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching. ASCD*.

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. National Council on Teacher Quality.

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.

Ferguson, R. F. (2015). The influence of teaching beyond standardized test scores: Engagement, mindsets, and agency. A study of 16,000. *Report from the achievement gap initiative*. Harvard University. *sixth through ninth grade classrooms*.

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–143). San Francisco: John Wiley & Sons, Inc.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica policy research reports*. https://EconPapers.repec.org/RePEc:mpr:mprres:0bbb1b46aa1c4c63b77dc9194c514e92.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality.

Grossman, P. (2015). *Protocol for Language Arts teaching observations (PLATO 5.0)*. Stanford University. http://platorubric.stanford.edu/index.html.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293–303. https://doi.org/10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470. https://doi.org/10.1086/669901

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D. & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. National Bureau of Economic Research, NBER Working Paper No. 16015.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., … Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113*(4), 461–487. https://doi.org/10.1086/669616

Hanushek, E. A. (2020). Education production functions. In S. Bradley, & C. Green (Eds.), *The economics of education* (2, pp. 161–170). Academic Press.

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

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher, 41*, 56–64. https://doi.org/10.3102/0013189X12437203

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching. Initial findings from the Measures of Effective Teaching project*. Bill & Melinda Gates Foundation.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives, 28*(62), 1–34. https://doi.org/10.14507/epaa.28.5012

Klette, K., Blikstad-Balas, M., & Roe, A. (2017). Linking Instruction and Student Achievement. A research design for a new generation of classroom studies. *Acta Didactica Norge, 11*(3), 1–19. https://doi.org/10.5617/adno.4729

Klette, K., Roe, A., & Blikstad-Balas, M. (2021). Observational scores as predictors for student achievement gains. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Why and how should we measure instructional quality* (pp. 173–203). Universitetsforlaget.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal, 48*(2), 361–386. https://doi.org/10.3102/0002831210382888

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice, 37*, 20–34. https://doi.org/10.1111/emip.12174

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the tripod student survey. *Educational Assessment, 22*(4), 253–274. https://doi.org/10.1080/10627197.2017.1381555

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction, 40*(2), 44–66.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*, 143–152. https://doi.org/10.1016/j.tate.2013.07.010

van der Lans, R. M. (2018). On the "association between two things": The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability, 30*(4), 347–366. https://doi.org/10.1007/s11092-018-9285-5

Lazarev, V., & Newman, D. (2014). Developing composite metrics of teaching practice for mediator analysis of program impact. *Empirical Education. Empowering educators for evidence-based decisions*.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist, 57*(9), 705–717. https://psycnet.apa.org/doi/10.1037/0003-066X.57.9.705.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-scale Assessments in Education, 6*(8), 1–11. https://doi.org/10.1186/s40536-018-0061-2

Luoto, J. M. (2023). *Comparative education and comparative classroom observation systems*. Comparative Education. https://doi.org/10.1080/03050068.2023.2173917

Luoto, J. M., Klette, K., & Blikstad-Balas, M. (2023). Possible biases in observation systems when applied across contexts: Conceptualizing, operationalizing, and sequenceing instructional quality. *Educational Assessement, Evaluation and Accountability, 35*, 105–128. https://doi.org/10.1177/17454999221077848

Maulana, M., & Helms-Lorenz, R. (2016). Observations and student perceptions of pre-service teachers' teaching behavior quality: Construct representation and predictive quality. *Learning Environments Research, 19*(3), 335–357. https://doi.org/10.1007/s10984-016-9215-8

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

Muijs, D., & Reynolds, D. (2018). *Effective teaching. Evidence and practice* (4th ed.). SAGE.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740–764. https://doi.org/10.1037/a0015576

Nilsen, T., & Gustafsson, J.-E. (Eds.). (2016), *Cohorts and time: 2. Teacher quality, instructional quality and student outcomes: Relationships across countries*. Springer.

OECD. (2018). *Education at a glance 2018: OECD indicators*. OECD Publishing. https://doi.org/10.1787/eag-2018-en. Reetrived from:.

OECD. (2020). *Global teaching InSights: A video study of teaching*. OECD Publishing. Reetrived from. https://www.oecd-ilibrary.org/education/global-teaching-insights_20d6f36b-en.

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*, 544–555. https://psycnet.apa.org/doi/10.1037/0022-0663.92.3.544.

Praetorius, A.-K., & Charalambous, C. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM, 50*, 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science, 24*(7), 1301–1308.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Scott, P., Mortimer, E., & Ametller, J. (2011). Pedagogical link-making: A fundamental aspect of teaching and learning scientific conceptual knowledge. *Studies in Science Education, 47*(1), 3–36. https://doi.org/10.1080/03057267.2011.549619

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

Senden, B., Nielsen, T., & Blömeke, S. (2021). Instructional quality: A review of conceptualizations, measurement approaches, and research findings. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of analyzing teaching quality: Potentials and pitfalls* (pp. 140–172). Universitetsforlaget.

Silseth, K., & Erstad, O. (2018). Connecting to the outside: Cultural resources teachers use when contextualizing instruction. *Learning, Culture and Social Interaction, 17*, 56–68. https://doi.org/10.1016/j.lcsi.2017.12.002

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis* (2 ed.). SAGE.

Spinath, B., & Steinmayr, R. (2012). The roles of competence beliefs and goal orientations for change in intrinsic motivation. *Journal of Educational Psychology, 104*(4), 1135–1148. https://psycnet.apa.org/doi/10.1037/a0028115.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Free Press.

Swedish Schools Inspectorate. (2012). Läsundervisning inom ämnet svenska för årskurs 7–9. Kvalitetsgranskning. [Reading instruction in language arts, school years 7–9. In *A quality examination. Rapport 2012:10. Swedish Schools Inspectorate*. Google Scholar.

Tengberg, M. (Ed.). (2022). *Undervisningskvalitet i svenska klassrum*. Studentlitteratur.

Tengberg, M., van Bommel, J., Nilsberth, M., Walkert, M., & Nissen, A. (2022). The quality of instruction in Swedish lower secondary language arts and mathematics. *Scandinavian Journal of Educational Research, 66*(5), 760–777. https://doi.org/10.1080/00313831.2021.1910564

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

White, M., & Klette, K. (2023). What's in a score? Problematizing interpretations of observation scores. *Studies In Educational Evaluation, 77*, Article 101238.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal and self-ratings in 360° feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179–192. https://doi.org/10.1023/A:1008158904681

Wilkinson, I. A. G., Murphy, P. K., & Binici, S. (2015). Dialogue-intensive pedagogies for promoting reading comprehension: What we know, what we need to know. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 37–50). American Educational Research Association.

Xu, L., & Clarke, D. (2018). Validity and comparability in cross-cultural video studies of classrooms. In L. Xu, G. Aranda, W. Widjaja, & D. Clarke (Eds.), *Video-based research in education: Cross-disciplinary perspectives* (pp. 19–33). Routledge.