More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science

Nani Teig[a,*], Ronny Scherer[b], Trude Nilsen[a]

[a] Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Norway
[b] Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Norway

ARTICLE INFO

ABSTRACT

Previous studies have assumed a linear relationship between inquiry-based teaching and student achievement in science. However this assumption may be questionable. Recent evidence on the effectiveness of inquiry-based teaching has yielded conflicting findings. To test the linearity assumption, the present study investigated the relation between inquiry-based teaching and achievement by taking into account the possible existence of nonlinear relations using Norwegian TIMSS 2015 data. A multilevel structural equation modeling analysis showed that the relationship was curvilinear. Inquiry-based teaching was positively correlated with achievement, but high frequency of inquiry activities was negatively related to achievement. Furthermore, we found that classroom SES did not affect the strength of the relationship between inquiry and achievement. These findings challenge the linearity assumption and contribute to explaining the conflicting evidence in earlier research as well as promote the effectiveness of inquiry-based teaching regardless of students' socioeconomic backgrounds.

Real science is a constant investigation of the unknown.
    – Abhijit Naskar, Neuroscientist, author, and speaker

1. Introduction

The relation between the implementation of certain instructional approaches and student achievement has probably become the most relevant indicator of teaching effectiveness. For instance, science education communities have long advocated the importance of inquiry-based teaching in improving student learning (e.g., American Association for the Advancement of Science, 1994; Osborne & Dillon, 2008; Rocard et al., 2007). The trend towards inquiry-based teaching can be seen as an attempt to develop students' reasoning and thinking skills through inquiry activities that represent the heart of scientific method (Klahr & Dunbar, 1988). This effort is partly reflected by the recommendations to implement inquiry-based methods to improve the quality of science teaching across European countries, such as the Rocard report *Science Education Now: A Renewed Pedagogy for the Future of Europe* (2007), and the increasing focus on experimentation and inquiry skills in educational large-scale assessments (e.g., Programme for International Student Assessment [PISA], Trends in International Mathematics and Science Study [TIMSS]). Despite these developments, an important question remains: Does inquiry-based teaching result in higher science achievement?

A growing body of research has investigated the effectiveness of inquiry teaching for improving student achievement. However, this research abounds in conflicting findings. Whereas some studies have documented a positive trend favouring inquiry-based instructional practices (e.g., Minner, Levy, & Century, 2010; Schroeder, Scott, Tolson, Huang, & Lee, 2007), the results of more recent studies using international large-scale assessment (ILSA) data indicated that this teaching strategy is negatively associated with science achievement (see Cairns & Areepattamannil, 2017). These inconsistent results in the literature might be due to several possible challenges associated with the nature of inquiry teaching measure and the analysis of the resultant data.

First, most studies investigating the effectiveness of inquiry-based teaching employ a measure of inquiry that focused on its frequency aspect (i.e., how often inquiry activities occur) and assume that a *linear* relationship exists between the inquiry activities and student achievement (e.g. Stohr-Hunt, 1996). Instead of assuming that "more is always better (or worse)", Creemers and Kyriakides (2008) argue that researchers investigating teaching effectiveness should consider the possibility of *nonlinear* associations of the variables under investigation. For instance, the implementation of inquiry-based science teaching to only a limited extent misses out on an important aspect of scientific literacy (NRC, 2013). At the same time, the implementation of this approach requires ample lesson time (Guskey, 2000), and an overemphasis on inquiry activities might cut into time spent on other necessary teaching and learning practices. Hence, it is possible to expect the presence of a curvilinear relation between inquiry-based teaching and student achievement.

---

* Corresponding author. Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Postbox 1099 Blindern, 0317 Oslo, Norway.
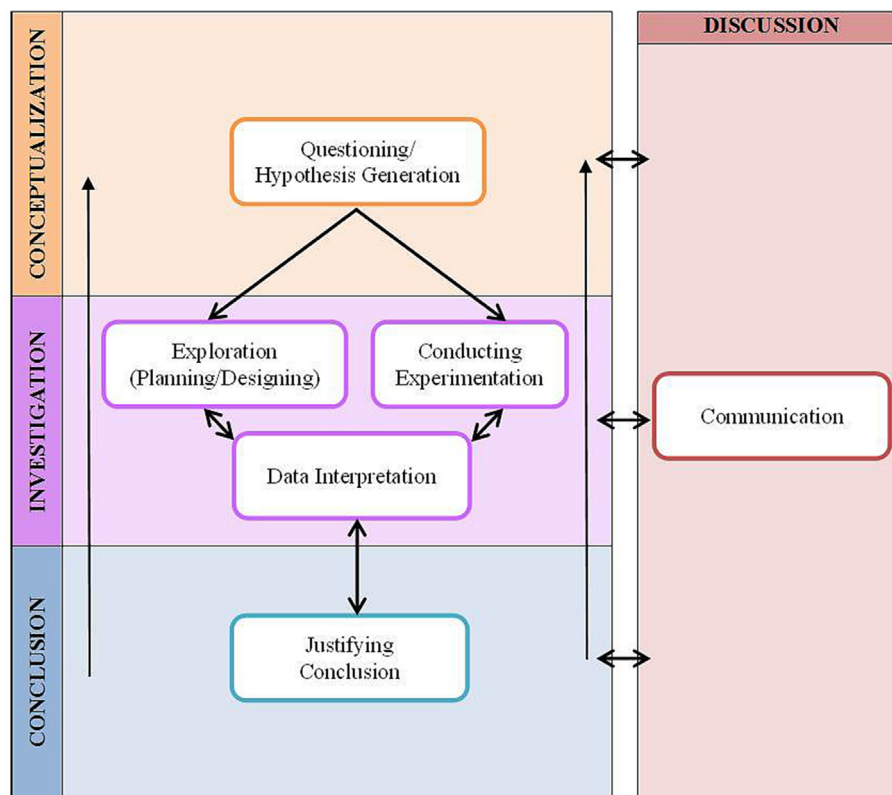*E-mail address:* nani.teig@ils.uio.no (N. Teig).

Fig. 1. The integrative phases of inquiry-based teaching activities in our study (a simplified inquiry-based learning framework from Pedaste et al., 2015, p. 56, p. 56).

Second, the use of ILSA data has attracted great attention in recent decades because it provides unique opportunities for generalizing findings to a wide population and examining teaching effectiveness across countries and cultures (e.g., Nilsen, Gustafsson, & Blömeke, 2016; Strietholt & Scherer, 2017). Since the implementation of inquiry-based teaching depends upon the teachers, variance in this construct may be more likely explained by differences between classes rather than between schools. Hence, to investigate the link between teaching strategies and student outcomes, the appropriate level of analysis—that is, the classroom level—is needed (Marsh et al., 2012). This presents a challenge for ILSA studies such as PISA, which primarily focuses on the student, school, and country levels rather than the classroom or teacher levels.

The aim of this study is therefore to shed light on the relationship between inquiry-based teaching and student achievement in science, based on a large-scale data set that overcomes the challenges associated with the *type of relationship* (i.e., linear vs. curvilinear) and *level of analysis* (i.e., classroom level vs. alternative levels). More specifically, we present findings from the TIMSS 2015—the only ILSA study that collects data from a nationally representative sample of schools and students in their intact science classrooms—and examine the relation between inquiry-based teaching activities and students' achievement, considering the possible existence of nonlinear relations and the appropriate unit of analysis (i.e., classrooms). In addition, it is important to examine whether and to what extent contextual variables, such as students' socioeconomic status (SES) in a class, might influence this relationship. As such, the effectiveness of inquiry instruction might vary by classroom SES. These findings contribute to an understanding of the role of inquiry-based teaching in science education and create awareness of the methodological challenges associated with the analysis of self-reported data, especially ILSA data for studies of teaching effectiveness in the research community.

1.1. Defining inquiry-based science teaching

Inquiry has been interpreted in many different ways across studies

and has become "one of the most confounding terms within science education" (Settlage, 2003, p. 34). Although there is a lack of agreement on the meaning of inquiry, it seems clear that this practice places a strong emphasis on promoting *active learning* and *student's responsibility for constructing knowledge* (de Jong & van Joolingen, 1998). Students are expected to acquire first-hand experience through inquiry practices in order to fully understand and appreciate how scientific knowledge is discovered (NRC, 2012).

In dealing with the non-uniform interpretations of inquiry, Abrams, Southerland, and Evans (2007) suggested focusing on the goals teachers have for applying inquiry in the classroom. In general, teachers implement inquiry practice so that their students can accomplish one or more of the following objectives: "to learn how to do science, learn about the nature of science, and learn science content"(NRC, 2000, p. 1). Based on the primary goal(s) of classroom inquiry, we define inquiry as the practice in which *students design or plan experiments, conduct experiments to collect evidence, interpret the evidence from the experiments, use the evidence to justify conclusions, and communicate the results of the experiments*. These activities frame inquiry as a student-centred approach by highlighting the importance of scientific investigations in achieving the goals of classroom inquiry.

We simplified the five phases of the inquiry-based learning framework developed by Pedaste et al. (2015) to illustrate various inquiry activities involved in our research. The original framework divided inquiry learning into five phases: Orientation, Conceptualization, Investigation, Conclusion, and Discussion. In our simplified phases of inquiry (see Fig. 1), we did not include the Orientation phase—which places more emphasis on teacher-centred approaches to stimulating students' interest in the scientific investigation at hand—since our definition of inquiry-based teaching views students as active learners who are responsible for their knowledge construction. Our adapted framework of inquiry starts with the Conceptualization phase, where students identify questions or formulate hypotheses to guide the inquiry process (Fig. 1). The research problems or hypotheses can be provided by the teacher, suggested by students, or identified by the teacher and students together. In the Investigation phase,

students make discoveries related to their questions by designing their investigations, conducting experiments, and interpreting and evaluating the outcomes. While the focus in the Exploration and Experimentation sub-phase is to collect reliable data, the Data Interpretation sub-phase places a strong emphasis on the process of meaning-making and building new knowledge from the data (Bruce & Casey, 2012). The next stage can be characterized as the Conclusion phase, which focuses on the process of comparing inferences drawn from the data to justify a conclusion. Students evaluate whether the research problems have been answered through the evidence collected from the investigation (Scanlon, Anastopoulou, Kerawalla, & Mulholland, 2011). Finally, the Discussion phase is viewed as a process that occurs during all phases. This phase represents an external process of inquiry and its openness, in which students discuss their findings and conclusions with the teacher and other students and receive feedback that can be used to improve the inquiry activities. Inquiry is viewed as an *integrated* and *nonlinear* process in which every activity is linked with each other in complex ways (Krajcik et al., 1998).

1.2. The relationship between inquiry-based teaching and student achievement in science

Inquiry has increasingly become a central term that represents "good science teaching and learning" in science education communities (Anderson, 2002). Researchers have advocated teachers to facilitate learners in constructing their own scientific knowledge and to engage them in the problem-solving process by using inquiry approach in their classrooms (Driver, Asoko, Leach, Scott, & Mortimer, 1994). Although it seems obvious that the implementation of scientific inquiry should be the basis for science teaching and learning, the evidence suggesting that inquiry-based teaching translates into increased student achievement, has been somewhat inconclusive.

In the past two decades, several meta-analyses and ILSA studies examining the effectiveness of inquiry-based science teaching have been reported. These studies can be divided into four major strands: *First,* studies that contrasted inquiry-based science teaching with other science teaching approaches highlighted the significance of inquiry approach in place of traditional passive teaching strategies (e.g., Minner et al., 2010; Schroeder et al., 2007). *Second,* empirical research that concentrated on the instructional guidance given to students found that activities involving teacher-guided instruction were more effective than independent or open-inquiry based approaches (Furtak, Seidel, Iverson, & Briggs, 2012; Jiang & McComas, 2015; Kang & Keinonen, 2017). However, the different types of guidance students received during scientific inquiry were equally effective in promoting learning (see Lazonder & Harmsen, 2016). *Third,* research that compared different types of inquiry activities across multiple countries showed that only the activities that emphasized models or applications in science and student-centred interactions were positively correlated with science achievement, and the opposite effect was found for inquiry that focused on hands-on scientific investigation activities (Areepattamannil, 2012; K. A. Gee & Wong, 2012; Lavonen & Laaksonen, 2009). *Fourth,* studies that considered student investigations to be the cornerstone of inquiry practice indicated that high levels of inquiry-oriented learning activities led to below-average levels of science performance (McConney, Oliver, Woods-McConney, Schibeci, & Maor, 2014). Similarly, recent research across 54 countries using PISA 2006 data revealed that inquiry-based science teaching was negatively related to students' scientific literacy (Cairns & Areepattamannil, 2017).

The preceding summary illustrates the variability in the ways that inquiry-based science teaching has been related to student achievement. Early research suggested that inquiry is a preferred instructional strategy in science classrooms. However, more recent studies—especially those using ILSA data places on the third and fourth strand of inquiry research—failed to show a positive correlation between scientific investigation activities and student achievement. We argue that this mixed evidence might have been caused by insufficient attention being paid to the *type of relationship* and *level of analysis* inherited in the ILSA data, which could have masked important differences in the effectiveness of inquiry-based science teaching across studies.

Regarding the *type of relationship*, a considerable amount of literature on teaching effectiveness has relied on the assumption that a linear association exists between teaching factors and student outcome variables. Most of these studies have examined the *frequency* dimension of teaching effectiveness, indicated by either the number of hours per week or other frequency responses (e.g., ranging from "never" to "every lesson"). Researchers have proposed the necessity of investigating nonlinear relations when examining the link between frequency dimension and student achievement (see Creemers & Kyriakides, 2008). For example, Caro, Lenkeit, and Kyriakides (2016) found that the association between cognitive activation strategies and mathematics achievement follows an inverted U-shape instead of linear pattern. Thus, the previous studies on inquiry effectiveness might have neglected the possible existence of a nonlinear relationship between inquiry-based teaching and achievement in science.

Choosing the appropriate *level of analysis* is another important consideration. In examining the impact of classroom-level constructs, the appropriate unit of analysis lies at the classroom level (for a comprehensive review see Marsh et al., 2012). Although many studies have stressed the need for measuring teaching effectiveness at the classroom level (for instance, Creemers & Kyriakides, 2008), several recent investigations failed to consider the importance of this analytical approach in studying the effect of inquiry-based science teaching on student achievement (e.g., Kang & Keinonen, 2017). This might be due to the fact that these studies were based on PISA data, which were collected at the student and the school level within a country. Even though investigating teaching effectiveness at the student or school level may provide insights into individual differences in student perceptions or the instructional climate in schools (Scherer, Nilsen, & Jansen, 2016), it suffers from methodological challenges associated with the inquiry effectiveness factors operating at the inappropriate level. To address this issue, we chose to use TIMSS data in the present study to examine inquiry-based science teaching at the classroom/teacher level.

While most studies on the relations between inquiry-based teaching and achievement would control for students' socioeconomic background—probably the most widely used students' contextual variable in teaching effectiveness research (Sirin, 2005)—few investigate the role SES plays in moderating this relation. The available literature suggests that students' access to educational resources at home is consistently related to higher science achievement, both at the student and classroom levels (e.g., Kaya & Rice, 2010). According to the *TIMSS 2015 International Results in Science* (Martin, Mullis, Foy, & Hooper, 2016), students and parents who reported owning many educational home resources (e.g., books at home) scored approximately 1.5 standard deviations higher on science achievement than those who reported having few resources. Although substantial evidence on the effects of SES on achievement exist, little is known about the role that SES plays in making a specific teaching approach—such as inquiry-based science teaching—successful.

The implementation of inquiry activities in classrooms with low average SES may be less fruitful in fostering student learning than in classes with high SES. As reported in several recent international studies, such as PISA 2015 (OECD, 2016) and TIMMS 2015 (Martin, Mullis, Foy, et al., 2016), students from low SES families are less likely to engage in academic activities. Interest towards science learning is a substantial factor for effective inquiry instruction since this practice often requires considerable effort and commitment from students to actively construct their own knowledge (Anderson, 2002). Furthermore, it is critical to note that students' engagement in inquiry activities, especially in the discussion phase, requires them to participate in a scientific discourse in which the use of language plays an important role in facilitating learning (Norris & Phillips, 2003; Wellington & Osborne, 2001). Research has shown that students who come from underprivileged SES families, especially those with immigrant backgrounds, are at a disadvantage when it comes to learning the language of science

(J. P. Gee, 2004; Ryoo, 2009). Along this line, it is valuable to explore whether the relation between inquiry-based teaching and student achievement is sensitive to SES differences. This line of thought adds a differential effectiveness perspective to the relation between inquiry activities and student achievement in science.

2. The present study

The present study utilizes a large dataset from TIMSS 2015 to investigate the relationship between inquiry-based teaching and science achievement by applying the appropriate level of analysis (i.e., the classroom level). First, we use a multilevel approach to investigate the possibility that inquiry-based teaching exhibits a curvilinear relationship with achievement. Second, we examine whether the association between inquiry-based teaching and science achievement changes after controlling for classroom SES and whether this association is moderated by the classroom SES. To our knowledge, this article is the first to report nonlinear relations between inquiry-based science teaching and achievement in addition to the effect of classroom SES.

3. Method

3.1. Sample and procedure

Data for the analyses were drawn from TIMSS, an international large-scale survey that assesses student performance in mathematics and science of four and eight graders in participating countries every fourth year (Martin, Mullis, Foy, et al., 2016). Our analyses were based on the Norwegian data from the TIMSS 2015 science grade 8th study. The sample consisted of $N = 4382$ students (50.1% girls) and 211 science teachers (52.7% female; 75.8% under the age of 50; teaching experience: $M = 13.1$ years, $SD = 10.4$ years). This study examined the data from the teacher and student background questionnaires in addition to the student science assessment.

TIMSS uses a two-stage stratified cluster sampling design in choosing participants within a country (Martin, Mullis, & Hooper, 2016). In the first stage, schools are sampled, and in the second stage, intact classrooms of students are selected at random within the participating schools. These classrooms are sampled instead of individual students across certain age groups or grade levels because TIMSS focuses on students' curricular and instructional experiences, which typically occur in classrooms. As such, students are nested within classrooms and classrooms within schools. To avoid bias associated with disproportional selection probabilities, sampling weights are computed at the school, classroom and student level, but only assigned at the school and classroom level since the student level weights are not informative (Kim, Anderson, & Keller, 2013; Rutkowski & Zhou, 2013; Stapleton, 2013).

3.2. Measures

3.2.1. Inquiry-based science teaching

Inquiry-based science teaching—represented as a latent variable—served as the independent variable in this study. Teachers were asked about their perceptions on the frequency of various activities in science lessons, "*In teaching science to the students in this class, how often do you ask them to do the following?*" They indicated the frequency with which 15 specific teaching and learning activities occurred in their science lessons using a 4-point Likert-type scale ranging from 1 (*never*) to 4 (*every or almost every lesson*). Of these 15 items, we chose the following five items that were related to our definition and phases of inquiry-based science teaching: (a) design or plan experiments or investigations [Expl], (b) conduct experiments or investigations [Expr], (c) present data from experiments or investigations [Com], (d) interpret data from experiments or investigations [Data], and (e) use evidence from experiments or investigations to support conclusions [Con].

3.2.2. Science achievement

We selected student achievement in science as a dependent variable in this study. The psychometric quality of the science assessment was evaluated extensively during the scaling procedure. For a detailed description of the process, we refer the reader to the Methods and Procedures in TIMSS 2015 report (Martin, Mullis, & Hooper, 2016). Student achievement was measured with a science test and estimated via a measurement model that produced a set of five plausible values to represent the range of student performance. All plausible values were incorporated into our models to produce an average of the model estimates (e.g., regression coefficients) and adjusted standard errors. Since our primary goal was to investigate the relationship between inquiry-based science teaching and student achievement at the classroom level, we used achievement scores at the classroom level and controlled for possible variation at the student level (Morin, Marsh, Nagengast, & Scalas, 2014).

3.2.3. Classroom SES

We used three items from the student background questionnaire (i.e., students' ratings of the number of books at home, the levels of their mother and fathers' highest education background) to create a latent variable of SES at the student level. This latent variable was also established at the classroom level under the doubly-latent variable framework to control for measurement error at the student and classroom level and sampling error in the aggregation of students' ratings to create SES construct at the classroom level (Morin et al., 2014).

3.3. Data analysis

Classroom level is the most appropriate level of analysis whenever research questions targeting teaching effectiveness are examined (Marsh et al., 2012). In the current data, 12.2% of variation in student achievement is explained at the classroom level. Accordingly, we performed multilevel structural equation modelling (MSEM) with students nested in classrooms. Our MSEM approach served two main purposes: (a) establishing measurement models to represent inquiry-based science teaching and student SES background; (b) examining the relations between teaching activities, SES, and science achievement. To accomplish (a), we used multilevel confirmatory factor analysis (MCFA)—an extension of CFA to multilevel situations that has proved useful to study the factor structure of instructional practices at two levels (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013)—inquiry-based science teaching and SES. Model fit was evaluated using Ryu's (2014) partial saturation approach by obtaining test statistics and fit indices for each level separately. Given that we used teacher reports to measure the frequency of inquiry-based activities in classrooms, the factor structure was only specified at the between level; the within level was empty. We referred to common guidelines for an acceptable model fit (i.e., $CFI \geq .95$, $TLI \geq 0.95$, $RMSEA \leq .08$, and $SRMR \leq .10$; Marsh, Hau, & Grayson, 2005). Nevertheless, we note that these guidelines do not represent "golden rules" (Marsh, Hau, & Wen, 2004), and the extent to which they apply to MSEM is still to be examined (Ryu, 2014). SES was based on student ratings that were aggregated to the classroom level, and the doubly-latent modelling approach was chosen to describe these contextual data (Morin et al., 2014).

We further performed MSEM on the basis of the previously examined factor structures. In these models, the relation between inquiry-based science teaching and achievement was described as linear or curvilinear. More specifically, we first specified a model assuming a linear relation among these constructs and evaluated its goodness-of-fit. In a second step, we added a quadratic component to examine whether a curvilinear relation described that data better. To compare the linear and curvilinear model, we performed the Satorra-Bentler corrected Likelihood-ratio test (Satorra & Bentler, 2010) and evaluated differences in information criteria (i.e., Akaike's Information Criterion [AIC], Bayesian Information Criterion [BIC], and sample-size adjusted BIC

[aBIC]) for each plausible-value dataset (Enders & Mansolf, 2016). We applied the latent moderated structural (LMS) equations approach to create the squared variable of inquiry-based science teaching and to analyse the latent interaction terms using the XWITH command in the software M*plus* (Muthén & Muthén, 1998–2015). We chose the LMS approach in our analyses to overcome the non-normality of the product term (Klein & Moosbrugger, 2000).

One of the complexities of the TIMSS 2015 data lies in the use of plausible values to represent science achievement (Martin, Mullis, Foy, et al., 2016). To account for this complexity, all analyses including science achievement were performed five times (i.e., once for each of the plausible-values datasets), and the resultant model parameters were pooled following Rubin's combination rules (Rubin, 1987). In this procedure, missing values that have occurred on the item level (less than 0.3% in the current data set) were handled by the expectation-maximization (EM) algorithm for each of the five plausible-value datasets. Robust maximum likelihood estimation (MLR) was used to address the non-normality and non-independence of the data. Item responses were treated continuously under the MLR estimator due to its robustness in the presence of categorical data with at least four response categories (Marsh et al., 2013; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

4. Results

4.1. Multilevel confirmatory factor analysis

We began by performing two MCFAs to investigate the fit indices for the measurement model of inquiry-based science teaching and student SES. For the inquiry-based science teaching construct, a single-factor model without any residual correlations indicated a poor fit, SB-χ^2(5) = 24.1, $p < 0.01$, RMSEA = 0.031, CFI = 0.903, TLI = 0.806, SRMR$_{within}$ = 0.000, SRMR$_{between}$ = 0.061, AIC = 1277, BIC = 1372, aBIC = 1324. In the next step, we added a residual correlation between the data interpretation and communication items for both methodological and conceptual reasons: the activities of interpreting data and discussing the results of an investigation are intertwined, and

modification indices suggested an improvement in model fit. The refined model in Fig. 2a showed an acceptable fit, SB-χ^2(4) = 4.1, $p > 0.05$, RMSEA = 0.002, CFI = 1.000, TLI = 0.999, SRMR$_{within}$ = 0.000, SRMR$_{between}$ = 0.025, AIC = 1248, BIC = 1349, aBIC = 1298. Standardized factor loadings were between 0.569 and 0.835, with a mean of 0.68 and a median of 0.66. Scale reliability was acceptable, $\omega = 0.75$.

For classroom-level SES, we followed the strategies outlined by Ryu and West (2009) for evaluating model fit separately at the individual and classroom level by specifying a one-factor model of SES at one level and a saturated model at the other level. We constrained the residual variance of the item "mothers' highest education level" at the classroom level to zero due to its insignificance. The model fit was satisfactory, both for the model with the SES factor structure at the student level and a saturated classroom-level as well as the corresponding model with a classroom-level factor structure of SES and a saturated-student level. These analyses did not indicate severe misfit, neither at the student nor the classroom level. We thus combined the two models by specifying the SES factor structure at both levels in a final step. The resultant, combined model in Fig. 2b fitted the data well, SB-χ^2(1) = 0.043, $p > 0.05$, RMSEA = 0.000, CFI = 1.000, TLI = 1.000, SRMR$_{within}$ = 0.000, SRMR$_{between}$ = 0.002, AIC = 32259, BIC = 32348, aBIC = 32304. The scale reliability at the classroom level was good, $\omega = 0.95$. Overall, the MCFA resulted in acceptable measurement models of inquiry-based science teaching and student SES.

4.2. Multilevel structural equation model

On the basis of the MCFA results, we evaluated a sequence of linear and curvilinear models to investigate the relation of inquiry to science achievement and the effect of classroom SES (see Fig. 3).

First, inquiry-based science teaching was not associated linearly with achievement (see Table 1; linear model 1.1). We then evaluated a curvilinear relationship by adding a quadratic term of inquiry and found a significant association with achievement (curvilinear model 2.1). The quadratic term of inquiry reflects the curvilinear relationship
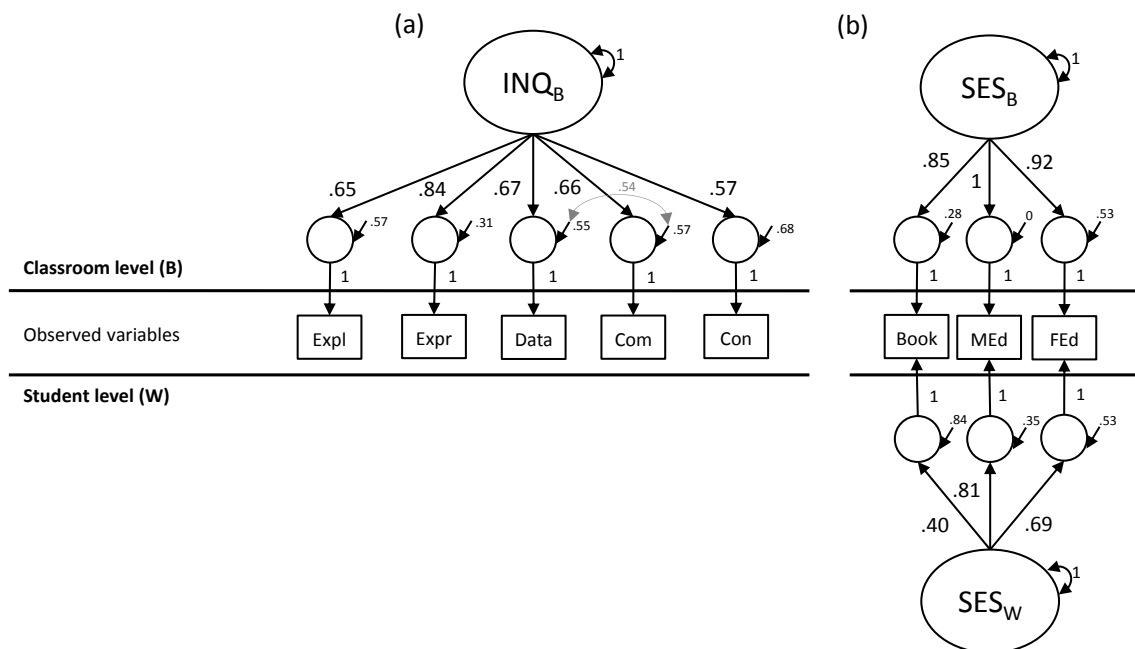

Fig. 2. Multilevel confirmatory factor analysis describing one-factor model of (a) inquiry-based science teaching and (b) SES at the classroom and student level.
*Note.* INQ = Inquiry-based science teaching, Expl = Exploration; design or plan experiments or investigations, Expr = Experiment; conduct experiments or investigations, Data = Data Interpretation; interpret data from experiments or investigations; Com = Communication; present data from experiments or investigations, Con = Conclusion; use evidence from experiments or investigations to support conclusions.
Book = number of books at home, MEd = mothers' highest education background, FEd = fathers' highest education background.
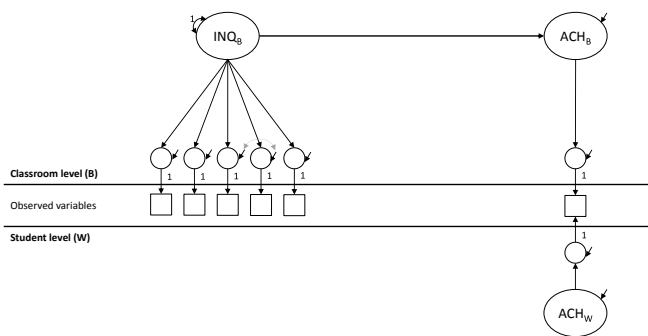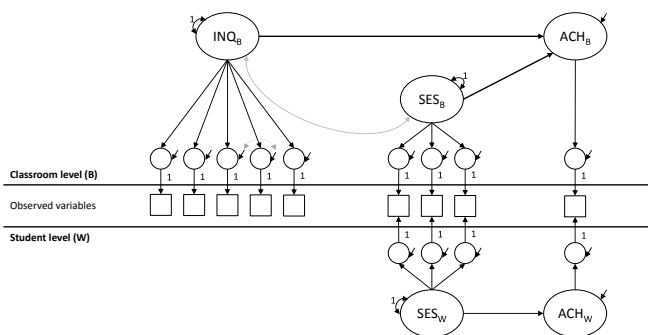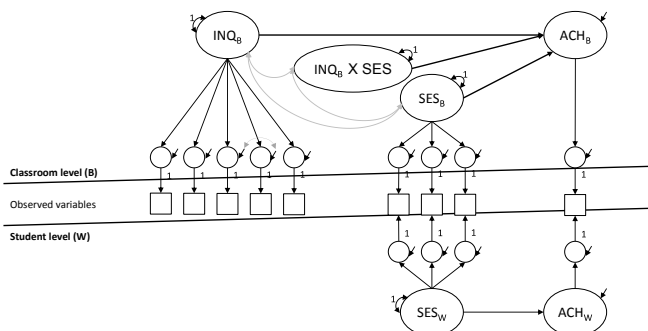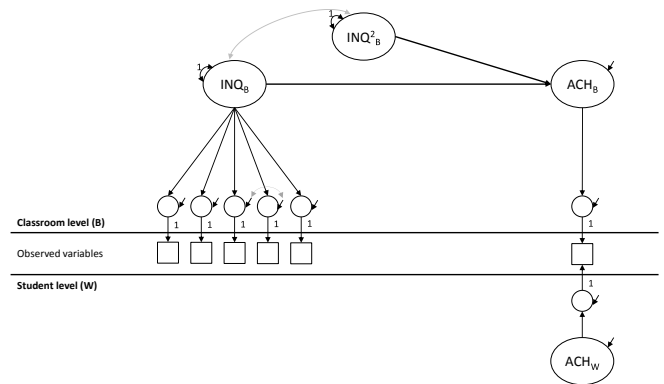
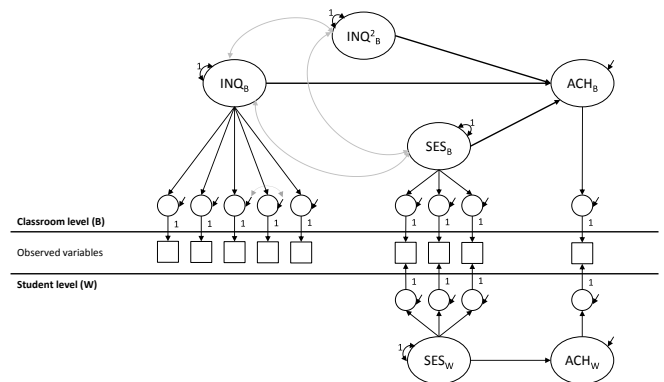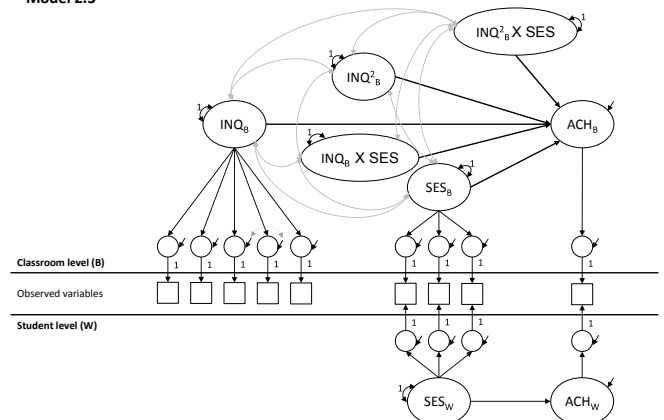**Fig. 3.** Multilevel structural equation model representing the linear and curvilinear relations among inquiry-based teaching and science achievement. Model 1.1 Linear relationships between inquiry-based teaching and science achievement. Model 1.2 Linear relationships between inquiry-based teaching and science achievement, controlling for SES. Model 1.3 Linear relationships between inquiry-based teaching and science achievement with the interaction effects of inquiry and SES. Model 2.1. Curvilinear relationships between inquiry-based teaching and science achievement. Model 2.2 Curvilinear relationships between inquiry-based teaching and science achievement, controlling for SES. Model 2.3 Linear relationships between inquiry-based teaching and science achievement with the interaction effects of inquiry and SES in addition to the interaction effects of inquiry-squared and SES. *Note.* ACH = Science achievement, INQ = Inquiry-based teaching, SES = Socioeconomic status. The item indicators of the quadratic term are not explicitly shown in the panel to sustain the clarity of presentation.

between the predictor and outcome, and the negative coefficient of the quadratic estimate indicates that the initially positive association between inquiry and achievement diminishes and becomes negative as the value of inquiry increases. Thus, this relationship takes the shape of an inverted U-curve rather than a straight line.

Second, using the Satorra-Bentler scaled likelihood ratio test, we compared the fit of the models assuming a linear and curvilinear relationship of inquiry-based science teaching and student achievement (see Table 2; linear model 1.1 and curvilinear model 2.1). The test statistics were significant for every plausible value of achievement, indicating that the curvilinear model fitted the data better than the

linear model (Satorra & Bentler, 2010).

Third, classroom SES had a strong and positive relationship with achievement in both models, which explained most of the variation in the outcome variable (see Table 1; linear model 1.2; curvilinear model 2.2). As a result of this strong relationship, the associations between inquiry-based science teaching and achievement decreased and became insignificant after controlling for SES at the student and the classroom level. Nevertheless, the results also showed that the associations between inquiry-based science teaching and achievement were not moderated by student SES, neither for linear nor for curvilinear models (linear model 1.3; curvilinear model 2.3).

Table 1

Latent linear and curvilinear regression of inquiry-based science teaching on science achievement.

Estimate	Linear model			Curvilinear model		
	Model 1.1	Model 1.2	Model 1.3	Model 2.1	Model 2.2	Model 2.3
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Regression coefficient[a]						
Inquiry	.072 (.096)	.072 (.071)	.066 (.069)	.122 (.082)	.098 (.067)	.097 (.067)
SES	–	.700 (.068)**	.640 (.060)**	–	.635 (.061)**	.643 (.078)**
Inquiry x Inquiry	–	–	–	-.264 (.130)*	-.174 (.116)	-.179 (.131)
Inquiry x SES	–	–	.055 (.162)	–	–	.022 (.182)
Inquiry^2 x SES	–	–	–	–	–	-.063 (.382)
R^2	.011	.678	.650	.031	.661	.667
Model fit[b]						
SB-χ^2 (*df*)	9.29 (8)	241.38 (27)	–	–	–	–
RMSEA	0.006	0.042	–	–	–	–
CFI	0.994	0.875	–	–	–	–
TLI	0.989	0.806	–	–	–	–
SRMR$_{within}$	0.000	0.066	–	–	–	–
SRMR$_{between}$	0.036	0.051	–	–	–	–
Npar	20	37	36	21	36	38
Log-likelihood (LL)	−5831.2	−21685.7	−21700.7	−5829.4	−21699.6	−21699.5
AIC	11702	43445	43473	11701	43471	43475
BIC	11831	43683	43705	11836	43703	43719
aBIC	11767	43566	43591	11769	43588	43599

Note. B = unstandardized regression coefficient; SB-χ^2 = Satorra-Bentler corrected χ^2 value (Satorra & Bentler, 2010), RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Root Mean Square Residual; Npar = number of free parameter; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; aBIC = sample-size adjusted BIC.

*p < .05 **p < .01.

[a] Science achievement scores are pooled from the five plausible values of achievement.

[b] Model fit indices using the LMS approach for the SB-χ^2, RMSEA, CFI, TLI, and SRMR are not yet provided by M*plus*.

The test statistics of model comparison for each plausible value of achievement showed a significant difference between the linear and curvilinear models, with inquiry and the classroom SES as the main effects (Table 2; linear model 1.2 and curvilinear model 2.2). In contrast, no significant difference was found on the models with the interaction effect of the classroom SES on the relations between inquiry and achievement in science (Table 2; linear model 1.3 and curvilinear model 2.3).

Overall, MSEM indicated that (a) a curvilinear relation between inquiry-based teaching and student science achievement existed and (b)

this relation was not moderated by classroom SES.

5. Discussion

The current study explored the link between inquiry-based teaching and student achievement in science. Our investigation extended previous research in two ways. First, it addressed the existence of a curvilinear relationship between the two constructs using relevant analysis of Norwegian TIMSS 2015 data at the classroom level; and second, it examined the effects of the student SES in a classroom from a

Table 2

The results of the Satorra-Bentler scaled likelihood ratio test for comparison between linear and curvilinear model for each plausible value of science achievement.

Model												Model comparisons				
Linear						Curvilinear						Plausible Value (PV)				
1.1		1.2		1.3		2.1		2.2		2.3		PV1	PV2	PV3	PV4	PV5
LL	SCF	LL	SCF	LL	SCF	LL	SCF	LL	SCF	LL	SCF	D(Δ*df*)	D(Δ*df*)	D(Δ*df*)	D(Δ*df*)	D(Δ*df*)
−5880	2.06					−5878	1.99					3.8 (1)*				
−5854	2.06					−5852	2.01						3.3 (1)*			
−5787	2.06					−5785	1.98							3.5 (1)**		
−5869	2.04					−5867	1.97								3.0 (1)*	
−5767	2.06					−5765	1.99									4.0 (1)*
		−21736	1.82					−21750	1.89			−26.5 (1)**				
		−21701	1.82					−21715	1.89				−27.1 (1)**			
		−21657	1.82					−21672	1.89					−29.0 (1)**		
		−21721	1.83					−21735	1.90						−29.2 (1)**	
		−21613	1.83					−21626	1.88							−26.5 (1)**
				−21751	1.89					−21750	1.85	2.8 (2)				
				−21715	1.88					−21714	1.83		1.5 (2)			
				−21673	1.88					−21672	1.85			2.3 (2)		
				−21736	1.89					−21735	1.84				2.1 (2)	
				−21628	1.88					−21626	1.83					2.8 (2)

Note. LL = Log-likelihood value; SCF = scaling correction factor for multilevel modelling; *D* = Difference in deviances (Δ[-2*LL]), the test statistics for the log-likelihood ratio test.

*p < .05 **p < .01.

differential effectiveness perspective.

5.1. The curvilinear relationship between inquiry-based teaching and student achievement in science

The implementation of inquiry-based teaching was related to science achievement. However, a linear relationship may not accurately reflect the nature of the association, as our findings indicated a curvilinear pattern (Model 2.1). The linear coefficient of inquiry was positive and non-significant, but the curvilinear coefficient was negative and significant, which indicates the presence of a curvilinear association between inquiry approach and student achievement. The increasing use of inquiry-based teaching was correlated with higher achievement in science until it reached an optimum value, then this association decreased as the use of the strategy increased. Furthermore, when there was a very high frequency of inquiry activities, the relationship between inquiry and achievement became negative. This nonlinear pattern indicates that excessive use of inquiry strategy in science classrooms may have diminishing returns for increasing student achievement because the frequency of student investigations per se does not lead to better performance. However, how the investigations are performed and discussed are of key significance for a more elaborate understanding of science concepts and principles (Duit & Tesch, 2010). As such, enacting inquiry activities in science classrooms is beneficial for student learning as it places a strong emphasis on the learners' development of scientific thinking (Dewey, 1910). As recommended by Lederman, Lederman, & Antink, "students will best learn scientific concepts by doing science" (2013, pp. 142–143).

Nevertheless, an overemphasis of inquiry-based instruction might disregard other necessary teaching strategies. For example, direct instruction was a more effective strategy in teaching students about experimental design than was discovery learning in a study Chen and Klahr (1999) conducted. Kirschner, Sweller, and Clark (2006) supported this observation and further argued that minimal guidance in discovery, problem-based, experiential, and inquiry-based teaching is probably the least effective teaching strategy compared to guided instruction. Using cognitive load theory, they illustrated how inquiry-based teaching contradicts the structure of human cognition by placing a heavy cognitive load on learner working memory and makes the learning process less effective. Furthermore, Stull and Mayer (2007) examined students' understanding of scientific texts by comparing the process of learning by doing and learning by viewing with graphic organizers. This study showed that students gained better understanding when they were provided with ready-made graphic organizers rather than constructing their own. They concluded that active learning that increases students' physical activity can create unnecessary cognitive load, which hinders student learning. Overall, these studies demonstrate that high emphasis on the frequent activities that involve student active learning, such as scientific investigations, may not always lead to deeper learning and underline our finding that "too much doesn't work". Excessive inquiry-based teaching would not be productive for student learning. Thus, complementing inquiry activities with direct instruction can help students develop a level of surface knowledge that is sufficient for effective deep learning processes (Hattie & Donoghue, 2016; Kirschner et al., 2006; Schneider & Preckel, 2017).

Another explanation for the limited effects from high frequencies of inquiry activities is related to the amount of resources needed to implement this instruction. Successful inquiry learning requires considerable time and efforts, both for teachers to plan an elaborate, well-thought lesson and for students to pursue a variety of inquiry activities (Guskey, 2000). Due to the resource-consuming nature of inquiry, teachers may encounter significant challenges as they try to prepare and manage the activities effectively, which can affect the quality of their instruction. Even when students are generally capable of conducting inquiry, Krajcik et al. (1998) have noted that they still face difficulties in recognizing the scientific value of the questions they

explore, collecting and analysing evidence systematically, and justifying their conclusions with appropriate data. Krajcik et al. further suggested that it would not be reasonable to expect students' inquiry skills to develop in a uniform manner, which highlights the role of instructional scaffolding in a successful inquiry practice. Although it depends on the complexity of the investigation addressed during the varying phases of inquiry, the extreme use of inquiry practice might negatively influence student achievement. Note that our findings should not be interpreted as encouraging a reduction in the use of inquiry-based science teaching. Inquiry instruction certainly plays a role in the development of students' scientific knowledge, which leads to increased achievement. As such, a high quality, rather than quantity, of inquiry activities is likely to be more beneficial to promote student learning (Marshall, Smart, & Horton, 2010).

While there have been a number of studies on the relations between inquiry-based teaching and achievement in science (e.g., Cairns & Areepattamannil, 2017), few recent studies have examined this association at the classroom level. For example, Gao (2014) conducted classroom level analyses and revealed that low, medium, and high achieving students were associated differently with inquiry-based teaching practice. In a similar study, Kaya and Rice (2010) found that inquiry was positively related to achievement in Singapore, but not in the USA and Australia. Even though these studies have applied the appropriate level of analysis, they did not take into account the frequency dimension of inquiry and only investigated the linear relationships between both constructs.

Due to the disparate and multiple interpretations of inquiry across studies that have resulted in confusion among teachers and scholars (Barrow, 2006), it is necessary to provide a clear definition of inquiry-based science teaching. In doing so, our study places a strong emphasis on a specific operationalization of inquiry grounded in the Pedaste et al. (2015) inquiry-based learning framework to help clarify what we mean by inquiry-based science teaching. While the previous studies imply inquiry-based science instructions involving separate inquiry activities, in the present study, these activities were viewed as interconnected phases within an integrative inquiry process (see Fig. 1) and supported by the measurement model of inquiry that fits the data well (see Fig. 2a). Accordingly, the integrated process becomes a distinctive and important feature of inquiry-based science teaching in this study.

By comparing the various conceptualizations of inquiry found across studies, Rönnebeck, Bernholt, and Ropohl (2016) distinguish two features of inquiry: the type and range of activities and the amount of teacher guidance. To address both features, Jiang and McComas (2015) analysed PISA 2006 data by using the inquiry framework from Schwab (1962) and Herron (1971) to classify students' inquiry activities into five levels based on the varying degree of guidance provided to them in designing investigations, conducting experiments, drawing conclusions, and asking questions. They found that Level 2 inquiry teaching, in which students fully engaged in conducting experiments and drawing conclusion, demonstrated the highest achievement. They also concluded that not all levels of inquiry are suitable and effective for all science instructions. Although we did not categorize inquiry activities based on the level of teacher guidance due to the difficulty in differentiating who is responsible for each activity in our data, this study indicates that frequency of activities is an important factor in examining the type and range of inquiry.

5.2. The effect of classroom SES

In line with previous studies (e.g., Kaya & Rice, 2010), it is evident that classroom SES predicts science achievement. Students in classes with a higher SES tend to have better achievement scores compare to those in a lower classroom SES. Moreover, even though Norway is known to have high levels of equity and small differences between schools (Martin, Mullis, Foy, et al., 2016; OECD, 2016), the relation between classroom SES and achievement ($B = 0.700$) seemed to be more comparable to that found internationally (e.g., Sirin, 2005). The strong relation between classroom

SES and achievement also leads to a reduced and insignificant effect of inquiry-based science teaching on achievement, as shown in Model 2.2. One explanation for this could be that SES alone explained 66.6% of the variation in achievement at the classroom level while inquiry only explained 3.1% (Model 2.1). Hence, the effect of SES on the outcome variable is so large that adding SES into any model would diminish the inquiry-achievement relation. It should be noted, however, that this should not merely be interpreted to mean that inquiry does not correlate with achievement after controlling for SES. According to the TIMSS 2015 International Results in Science, the index of teacher emphasis on inquiry investigation on about half the lessons or more was 5% in Norwegian eighth grade classrooms (Martin, Mullis, Foy, et al., 2016). As such, variation of the frequency of inquiry activities is very small, which can lead to a low statistical power. Furthermore, although the individual sample size is quite large ($N = 4382$ students), these students are nested in 211 classrooms. In a multilevel analysis, a large number of classrooms are more important than the student-level sample size (Maas & Hox, 2005). The sample size of 211 classrooms may have not been large enough and could therefore also contribute to the lack of power in addition to the low frequency of inquiry investigation in Norway.

Despite these findings, there was no evidence that students from high- and low-SES classrooms benefit differently from inquiry-based science teaching. Inquiry practice seems to be beneficial for students' achievements regardless the levels of their classrooms SES. Contrary to Secker's findings (2002) that inquiry-based teacher practices contribute to larger achievement gaps between high and low advantaged students, the results of this study indicate that the inquiry-achievement relation was not influenced by varying levels of students' economic backgrounds in a classroom. This instructional approach "appears to make science interesting and accessible to a wider range of students than is possible with traditional approaches" (White & Frederiksen, 1998, p. 91). Our study therefore contributes to the growing body of evidence supporting the advocacy of enacting inquiry-based science teaching, including across classrooms with diverse economic backgrounds (e.g., Blanchard et al., 2010).

5.3. Limitations and future directions

Since the results of this study were based on the secondary analysis of TIMSS data, there are some important limitations for research on teaching effectiveness. First, the TIMSS measure of inquiry-based science teaching was based on teacher ratings. Adding students' perspectives or researchers' classroom observations to these ratings may substantiate our findings even further. Second, given the cross-sectional nature of the TIMSS data, inferences about the causality of the inquiry-achievement relation cannot be drawn. However, the results obtained from this study should stimulate experimental investigations of the curvilinear relationship. Third, most of the items that were used to measure inquiry did not specify the extent of guidance teachers provided in varying inquiry activities. Providing a clear description of who is responsible for each activity could contribute to making the conceptualization of inquiry-based science teaching in TIMSS questionnaire more comprehensive. Fourth, our study focused on the analysis of teachers' perceptions of the *frequency* rather than the *quality* of inquiry activities. Additionally, we measured teacher beliefs about the frequency of inquiry activities instead of actual frequency. Taking a more qualitative perspective on these activities, such as through video observations, could provide deeper insights into the optimal quantity and quality of inquiry activities. Although this kind of research is generally costly and difficult to standardize across classrooms, it captures important characteristics of the inquiry approach as it is enacted in actual classrooms.

5.4. Conclusion

This research investigates the relationship between inquiry-based teaching and student achievement in science as well as the effect of the classroom SES in curvilinear models. Our findings imply that inquiry-based science teaching is positively related to students' science performance. We extend earlier research by examining the fading effects of this teaching strategy on achievement, especially with high-frequency inquiry activities. Our results provide insight into the multifaceted nature of the association between inquiry and achievement by demonstrating the existence of a curvilinear pattern. This challenge the assumption of linearity—that "more is always better (or worse)"—and contributes to explaining the conflicting evidence found in previous empirical research. Our study also confirms that varying levels of classroom SES are associated with student achievement; yet our results show it does not influence the strength of the association between inquiry activities and achievement in science. Thus, these results further promote the effectiveness of inquiry-based science teaching to foster learning in all students, regardless of their socioeconomic backgrounds.

In conclusion, our findings are particularly useful for stimulating future dialog among researchers, policymakers, and practitioners on the role of inquiry-based teaching in science education. This study shall also create an awareness of the methodological challenges in describing instructional effectiveness using ILSA data.

Acknowledgement

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.learninstruc.2018.02.006.

References

Abrams, E., Southerland, S. A., & Evans, C. (2007). Inquiry in the classroom: Necessary components of a useful definition. In E. Abrams, S. A. Southerland, & P. Silva (Eds.). *Inquiry in the classroom: Realities and opportunities*. Charlotte, North Carolina: Information Age Publishing, Inc.

American Association for the Advancement of Science (1994). *Science for all americans: Project 2061*. Oxford: Oxford University Press.

Anderson, R. D. (2002). Reforming science Teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1–12.

Areepattamannil, S. (2012). Effects of inquiry-based science instruction on science achievement and interest in Science: Evidence from Qatar. *The Journal of Educational Research, 105*(2), 134–146.

Barrow, L. H. (2006). A brief history of inquiry: From Dewey to standards. *Journal of Science Teacher Education, 17*(3), 265–278.

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability?: A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education, 94*(4), 577–616.

Bruce, B. C., & Casey, L. (2012). The practice of inquiry: A pedagogical 'sweet spot' for digital literacy? *Computers in the Schools, 29*(1–2), 191–206.

Cairns, D., & Areepattamannil, S. (2017). Exploring the relations of inquiry-based teaching to science achievement and dispositions in 54 countries. *Research in Science Education,* 1–23.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies In Educational Evaluation, 49*, 30–41.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

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

Dewey, J. (1910). Science as subject-matter and as method. *Science, 31*(787), 121–127.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5–12.

Duit, R., & Tesch, M. (2010). On the role of the experiment in science teaching and learning–Visions and the reality of instructional practice. In M. Kalogiannakis, D. Stavrou, & P. Michaelides (Eds.). *Proceedings of the 7th international conference on hands-on science* (pp. 17–30). Rethymno: The University of Crete.

Enders, C. K., & Mansolf, M. (2016). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*. http://dx.doi.org/10.1037/met0000102 No Pagination Specified.

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.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Review of Educational*

Research, 82(3), 300–329.

Gao, S. (2014). Relationship between science teaching practices and students' achievement in Singapore, Chinese Taipei, and the US: An analysis using TIMSS 2011 data. *Frontiers of Education in China, 9*(4), 519–551.

Gee, J. P. (2004). Language in the science classroom: Academic social languages as the heart of school-based literacy. In R. K. Yerrick, & W.-M. Roth (Eds.). *Establishing scientific classroom discourse communities: Multiple voices of teaching and learning research* (pp. 19–37). Mahwah, NJ: Lawrence Erlbaum Associates.

Gee, K. A., & Wong, K. K. (2012). A cross national examination of inquiry and its relationship to student performance in science: Evidence from the Program for International Student Assessment (PISA) 2006. *International Journal of Educational Research, 53*, 303–318.

Guskey, T. R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin Press.

Hattie, J. A. C., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *Npj Science Of Learning, 1*, 16013.

Herron, M. D. (1971). The nature of scientific enquiry. *The School Review, 79*(2), 171–212.

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and Attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education,* 1–23.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179–201.

Kang, J., & Keinonen, T. (2017). *The effect of student-centered approaches on students' interest and achievement in Science: Relevant topic-based, open and guided inquiry-based, and discussion-based approaches. Research in science education.*

Kaya, S., & Rice, D. C. (2010). Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education, 32*(10), 1337–1363.

Kim, J.-S., Anderson, C. J., & Keller, B. (2013). Multilevel analysis of assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.). *Handbook of international large-scale Assessment: Background, technical issues, and methods of data analysis* (pp. 389–425). Boca Raton, FL: CRC Press.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457–474.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science Classrooms: Initial attempts by middle school students. *The Journal of the Learning Sciences, 7*(3/4), 313–350.

Lavonen, J., & Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching, 46*(8), 922–944.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681–718.

Lederman, N. G., Lederman, J. S., & Antink, A. (2013). Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy. *International Journal of Education in Mathematics, Science and Technology, 1*(3).

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

Marshall, J. C., Smart, J., & Horton, R. M. (2010). The design and validation of equip: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education, 8*(2), 299–321.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares, & J. J. McArdle (Eds.). *Contemporary psychometrics* (vol. 275-340). Mahwah, NJ: Lawrence Erlbaum.

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: A Multidisciplinary Journal, 11*(3), 320–341.

Marsh, H. W., Ludtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual Effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124.

Marsh, H. W., Vallerand, R. J., Lafreniere, M. A., Parker, P., Morin, A. J., Carbonneau, N., et al. (2013). Passion: Does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychological Assessment, 25*(3), 796–809.

Martin, M., Mullis, I., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in science.* Boston, MA: Boston College, TIMSS & PIRLS International Study Center.

Martin, M., Mullis, I., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

McConney, A., Oliver, M. C., Woods-McConney, A., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education, 98*(6), 963–980.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction-what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496. http://dx.doi.org/10.1002/tea.20347.

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom Climate: An illustration. *The Journal of Experimental Education, 82*(2), 143–167.

Muthén, B., & Muthén, L. (1998-2015). *Mplus (version 7.3).* Los Angeles, CA: Muthén & Muthén.

National Research Council (2000). *Inquiry and the national science education standards: A guide for teaching and learning.* Washington, DC: The National Academies Press.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press.

National Research Council (2013). *Next generation science Standards: For states, by states.* Washington, DC: The National Academies Press.

Nilsen, T., Gustafsson, J.-E., & Blömeke, S. (2016). Conceptual framework and methodology of this report. In T. Nilsen, & J.-E. Gustafsson (Eds.). *Teacher quality, instructional quality and student Outcomes: Relationships across countries, cohorts and time* (pp. 1–19). Cham, Switzerland: Springer International Publishing.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224–240.

OECD (2016). *PISA 2015 results (volume I): Excellence and equity in education.* Paris: OECD Publishing.

Osborne, J. F., & Dillon, J. (2008). *Science education in Europe: Critical reflections, Vol. 13.* London: The Nuffield Foundation.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61.

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education Now: A renewed Pedagogy for the future of Europe. Retrieved from Brussels).*

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground–A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–197.

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

Rutkowski, L., & Zhou, Y. (2013). Using structural equation models to analyze ILSA data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.). *Handbook of international large-scale Assessment: Background, technical issues, and methods of data analysis* (pp. 425–450). Boca Raton, FL: CRC Press.

Ryoo, K. (2009). *Learning science, talking science: The impact of a technology-enhanced curriculum on students' science learning in linguistically diverse mainstream classrooms*Unpublished doctoral dissertation Stanford, USA: Stanford University. Retrieved from https://search.proquest.com/docview/305009167.

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology, 5.* http://dx.doi.org/10.3389/fpsyg.2014.00081.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601.

Satorra, A., & Bentler, P. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248.

Scanlon, E., Anastopoulou, S., Kerawalla, L., & Mulholland, P. (2011). How technology resources can be used to represent personal inquiry and support students' understanding of it across contexts. *Journal of Computer Assisted Learning, 27*(6), 516–529.

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional Quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7.*

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T. Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436–1460.

Schwab, J. J. (1962). The teaching of science as enquiry. In J. J. Schwab, & P. F. Brandwein (Vol. Eds.), *The teaching of science: Vol. 253.* Cambridge, MA: Harvard University Press.

Secker, C. v (2002). Effects of inquiry-based teacher practices on science excellence and equity. *The Journal of Educational Research, 95*(3), 151–160.

Settlage, J. (2003). Inquiry's allure and illusion: Why it remains just beyond our reach. *Paper presented at the annual meeting of the national association for research in science teaching, Philadelphia.*

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Stapleton, L. (2013). Incorporating sampling weights into single-and multilevel analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.). *Handbook of international large-scale Assessment: Background, technical issues, and methods of data analysis* (pp. 363–388). Boca Raton, FL: CRC Press.

Stohr-Hunt, P. M. (1996). An analysis of frequency of hands-on experience and science achievement. *Journal of Research in Science Teaching, 33*(1), 101–109.

Strietholt, R., & Scherer, R. (2017). The contribution of international large-scale assessments to educational Research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research,* 1–18.

Stull, A. T., & Mayer, R. E. (2007). Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology, 99*(4), 808.

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.

Wellington, J., & Osborne, J. (2001). *Language and literacy in science education.* Buckingham, UK: McGraw-Hill Education.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and Metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.