Manuscript Details

Manuscript number	JLI_2018_328_R2
Title	The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England.
Article type	Research Paper

Abstract

Inquiry-based science teaching involves supporting pupils to acquire scientific knowledge indirectly by conducting their own scientific experiments, rather than receiving scientific knowledge directly from teachers. This approach to instruction is widely used among science educators in many countries. However, researchers and policymakers have recently called the effectiveness of inquiry approaches into doubt. Using nationally-representative, linked survey and administrative data, we find little evidence that the frequency of inquiry-based instruction is positively associated with teenagers' performance in science examinations. This finding is robust to the use of different measures of inquiry, different examinations/measures of attainment, across classrooms with varying levels of disciplinary standards and across gender and prior attainment subgroups.

Keywords	PISA; science; inquiry-based instruction.

Corresponding Author	John Jerrim
Corresponding Author's Institution	UCL
Order of Authors	John Jerrim, Mary Oliver, Sam Sims

Submission Files Included in this PDF

File Name [File Type]

Cover_Letter.docx [Cover Letter]

Referee_Responses_22_11_2018.docx [Response to Reviewers (without Author Details)]

Highlights22_11_18.docx [Highlights]

Title_Page_22_11_2018.docx [Title Page (with Author Details)]

Main_Body_Revision_22_11_2018.docx [Manuscript (without Author Details)]

Online_Supplementary_Material_22_11_18.docx [MethodsX]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England.

John Jerrim (UCL Institute of Education and Education Datalab)

Mary Oliver (University of Nottingham)

Sam Sims (UCL Institute of Education and Education Datalab)

November 2018

Inquiry-based science teaching involves supporting pupils to acquire scientific knowledge indirectly by conducting their own scientific experiments, rather than receiving scientific knowledge directly from teachers. This approach to instruction is widely used among science educators in many countries. However, researchers and policymakers have recently called the effectiveness of inquiry approaches into doubt. Using nationally-representative, linked survey and administrative data, we find little evidence that the frequency of inquiry-based instruction is positively associated with teenagers' performance in science examinations. This finding is robust to the use of different measures of inquiry, different examinations/measures of attainment, across classrooms with varying levels of disciplinary standards and across gender and prior attainment subgroups.

Key Words: PISA, science, enquiry-based instruction.

Contact details: John Jerrim (J.Jerrim@ucl.ac.uk) Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL

Acknowledgements: This research has been funded by the Nuffield Foundation. We are grateful for their support.

The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England.

Inquiry-based science teaching involves supporting pupils to acquire scientific knowledge indirectly by conducting their own scientific experiments, rather than receiving scientific knowledge directly from teachers. This approach to instruction is widely used among science educators in many countries. However, researchers and policymakers have recently called the effectiveness of inquiry approaches into doubt. Using nationally-representative linked survey and administrative data we find little evidence that the frequency of inquiry-based instruction is positively associated with teenagers' performance in science examinations. This finding is robust to the use of different measures of inquiry, different measures of attainment, across classrooms with varying levels of disciplinary standards and across gender and prior attainment subgroups.

Key Words: PISA, science, inquiry-based instruction.

Introduction

It has long been recognised that science skills are important for technological innovation (Varsakelis, 2006) and economic growth (Hanushek & Woessmann, 2012). Many have argued that scientifically literate young people are also better equipped to make choices and decisions that impact their lives and environment (e.g. Rutherford & Ahlgren, 1991). Governments across the world are consequently looking for the most effective ways to improve scientific education. Science teachers clearly play a pivotal role in developing the next generation of scientists. However, they face a critical question – what is the most effective way to teach science?

One prominent school of thought is that science should be taught using inquiry methods. At a high level, inquiry is an active form of learning (Sjøberg, 2015) which involves pupils answering research questions using data (Bell, Smetana & Binns, 2005). Inquiry teaching aims to provide students with knowledge via investigation, rather than receiving knowledge directly from teachers (Lazonder & Harmsen, 2016). A more granular description of inquiry (Pedaste et al., 2015) breaks these activities down into phases: orientation (in which the topic is introduced and motivated); conceptualization (in which a research question and/or hypothesis is developed); investigation (in which observation and experiment are conducted and data interpreted); conclusion (in which inferences are drawn and models or hypotheses are evaluated); discussion (in which findings are communicated). Inquiry therefore incorporates teaching science process skills (teaching *of* inquiry), teaching how scientists use inquiry methods (teaching *about* inquiry) and teaching scientific knowledge using inquiry process skills (teaching *through* inquiry) (Cairns & Areepattamannil, 2017). Yet despite broad brush agreement on the processes and aims of inquiry teaching there remains considerable variability in the way that inquiry has been implemented and operationalised in the literature. Rönnebeck, Bernholt and Ropohl (2016) emphasise that different studies place different emphasis upon two important dimensions of inquiry: the types of activities pupils engage in and the degree of guidance provided by teachers. This can hinder the comparability and accumulation of results across studies.

In order to ensure the present study contributes to the existing literature we adopt a widespread and longstanding definition that incorporates all the aspects of inquiry set out in this paragraph. Specially, we adopt the 1996 National Science Education Standards (NSES) definition which states that inquiry teaching aims to help pupils:

'develop the ability to think and act in ways associated with inquiry, including asking questions, planning and conducting investigations, using appropriate tools and techniques to gather data, thinking critically and logically about relationships between evidence and explanations, constructing and analyzing alternative explanations, and communicating scientific arguments' (NRC, 1996, p. 105).

We adopt the NSES definition for three reasons. First, because this definition has been used in all five of the major reviews of inquiry-based teaching published since 1996 (Anderson, 2002; Minner, Levy, & Century, 2010; Furtak, Seidel, Iverson, & Briggs, 2012; Lazonder & Harmsen, 2016; Ronnebeck et al., 2016). Despite being dropped from later NRC standards documents (NRC, 2012), it has therefore become the standard definition in the academic literature. Second, we adopt this definition because the NSES (NRC, 1996; 2000) have directly informed the development of the questionnaires used to collect the PISA data used in the present study (OECD, 2006; OECD, 2016). This suggests that the OCED measures of inquiry teaching will therefore be aligned with the NSES definition. Third, because the NSES definition (or a very close equivalent) has been employed by science associations and government agencies in many countries outside the US, including in England (ASE, 2017) and Europe (Rocard et al., 2007).

A global movement for improving science education in schools using more inquiry-based approaches has been evident for several years (see Bell, Urhahne, Schanze, & Ploetzner, 2010; Furtak et al., 2012; Lazonder & Harmsen, 2016; Minner et al., 2010). For example, the European Union (EU) has funded several projects arguing that improvements in science education could be brought about with the introduction of inquiry-based approaches in schools (Rocard, 2007). In England, the setting for our empirical analysis, the inquiry approach retains the support of science teaching associations, influential science research funders and at least until recently, the national school inspectorate (ASE, 2009; Holman, 2017; Ofsted, 2013). Arguments for inquiry science are grounded in the constructivist belief that asking pupils to solve authentic problems and allowing them to construct their own solutions or distil their own understanding makes the learning experience more meaningful (Kirschner, 1992; Pressley et al., 2003). For example, Minner et al. (2010) claim that students learning through scientific investigations 'are more likely to increase conceptual understanding than are strategies that rely on more passive techniques' (Minner et al., 2010, p. 474). Others have also claimed that students are more motivated by this approach (e.g. <*blind for review*>).

The value of inquiry-based teaching is however strongly contested (Hodson, 2014; Kirschner, Sweller & Clark, 2006; Zhang, 2016). Critics of inquiry-based instruction argue that it overlooks important features of cognitive architecture (Kirschner et al., 2006; Rosenshine, 2012; Zhang, 2016). More specifically, they point to evidence that pupils' limited working memory is likely to be overloaded by the difficulty of conducting scientific investigations which may serve to limit rather than facilitate the acquisition of new knowledge. In England, the context for this study, Nick Gibb (the Minister for Schools) has publicly denounced the use of inquiry methods citing cross-sectional evidence from PISA to claim that they are ineffective:

"allowing pupils to design their own experiments; allowing pupils to investigate and test their ideas; holding class debates about investigations; and requiring pupils to argue about science questions…resulted in a net negative impact on science outcomes*"* (Gibb, 2017).

There is therefore still considerable debate about whether inquiry is the best method for teaching science.

Empirical researchers have tried to address the debate using data. Critics of inquiry-based methods point to the results from meta-analyses which have shown positive causal effects of direct instruction (in which teachers provide knowledge directly to students rather than helping them acquire knowledge through investigation) when compared to 'business as usual' in a range of subjects (Stockard, Wood, Coughlin, & Rasplica Khoury, 2018) and for science in particular (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011). Advocates of inquiry-based teaching have in turn responded by arguing that the evidence for direct instruction is strongest in mathematics and point to science-specific meta-analyses that have found inquiry-based methods to be more effective (Furtak et al., 2012). Moreover, inquiry-based teaching advocates argue that poorly designed or executed forms of inquiry-based instruction are often used as the counterfactual in studies that directly compare one form of instruction to the other. The present research aims to contribute to this debate by answering the following two research questions.

Research Question 1: *Do young people who receive a higher frequency of inquiry-based science teaching have higher levels of science achievement?*

Research Question 2: *Is there a positive association between specific components of inquiry-based teaching and young people's achievement in science?*

Some researchers have argued that inquiry-based teaching is most effective when it involves higher levels of guidance (e.g. Hmelo-Silver, Duncan, & Chin, 2007). To see how it is

conceptually meaningful to have guided inquiry recall that Rönnebeck et al. (2016) identified two dimensions of inquiry in the literature: activities undertaken by students and level of guidance. Guided inquiry therefore maintains the emphasis on acquiring knowledge indirectly through investigations conducted by pupils (the activities dimension) but increases the level of guidance provided by the teacher (the guidance dimension). While this clearly constitutes a departure from pure inquiry-based teaching, even the original proponents of this approach admitted that there could be variations in the extent to which inquiry was guided (Schwab, 1962).

The justification for guided inquiry (Martin, 2016) derives from cognitive load theory (CLT) (Sweller, Ayres, & Kalyuga, 2011a). CLT states that learners must process new information in their working memory. Working memory is however limited in its capacity and if this is overloaded, learning will be impeded. The cognitive load involved in learning new material can be divided into intrinsic load, which is inherent to the information being learned, and extraneous load, which is contingent on the instructional methods through which the information is acquired (Sweller, Ayres, & Kalyuga, 2011b). Guided inquiry therefore aims to minimise extraneous load (Martin, 2016) either by placing constraints around the inquiry being conducted or providing cues and prompts to aid students in the process of inquiry (De Jong & Lazonder, 2014). Meta-analytic evidence suggests that more guided forms of inquiry learning tend to be more effective than unguided discovery in science (Furtak et al., 2012; Lazonder & Harmsen, 2016). We provide new evidence on this through our third and final research question:

Research Question 3: *Are guided approaches to inquiry instruction positively associated with pupils' achievement in science?*

This paper adds value to the existing literature in three ways. First, a significant limitation of existing studies on inquiry teaching using the Programme for International Student Assessment (PISA) and Trends in Mathematics and Science Study (TIMSS) datasets is that they are based upon cross-sectional data and do not control for measures of prior achievement. In contrast, our linked data includes prior attainment measures from high-stakes, externally-marked examinations conducted just prior to entering secondary school. Second, much of the existing literature used tests administered specifically for the purposes of the study. Our linked data allows us to use both high-stakes, externally-marked examinations conducted at the end of secondary school and PISA tests, which are well aligned with the aims of inquiry-based

teaching. Third, many of the existing experimental studies use small samples or involve laboratory experiments with limited ecological validity. By contrast, our study draws upon rich, nationally-representative data including more than 4,000 15/16-year-olds. In summary, our study is one of very few able to investigate the effects of inquiry instruction as implemented by teachers in natural school settings using rich, longitudinal data including high-stakes measures of science attainment.

Data

The PISA sample design

PISA is an international study of 15-year-olds' academic achievement. Rather than attempting to assess pupils' knowledge of national curricula, PISA attempts to capture how well young people can apply reading, science and mathematics skills in real-world situations. We use the data for England from the most recent cycle (2015), when science was the subject of focus. A two-stage sample design was used. Schools were first sampled with probability proportional to size and then pupils were randomly selected from within schools. A total of 5,194 pupils from 206 schools in England participated in PISA 2015. This amounts to a 92 percent response rate at the school level and an 88 percent response rate at the pupil level[1]. In England almost all participating pupils are within the same year group (Year 11).

The PISA 2015 sample for England has been linked to the National Pupil Database (NPD), which includes administrative data upon pupils' backgrounds along with their performance on national examinations. A successful link has been made between PISA and the NPD for 95 percent of the full sample[2]. The NPD link also provides us with access to other prior cognitive achievement measures such as pupils' Key Stage 2 (KS2) scores at age 11 in science (teacher-assessed), reading and mathematics.

[1] School-level response rates in England were 83 percent before replacement schools were included and 92 percent after.
[2] Independent school pupils were less likely to have linked GCSE data than state school pupils. The high overall linkage rate should mean that this has only a relatively minor impact upon our results.

Measures of science achievement

Our primary outcome measure is pupils' grades in their science General Certificate of Secondary Education (GCSE) examination. This is a high-stakes, externally-marked exam taken by all pupils at age 16. We focus upon children's science pillar score because it maximises the amount of comparable information on GCSE science performance for our sample[3]. It is important to note that the Year 11 pupils in our sample first took the PISA science assessments in November/December 2015 and then sat their GCSEs in May/June 2016 – just six months apart. The strength of the GCSE outcome measure is that it is based upon children's achievement in a high-stakes examination. Issues around low-test motivation, marking or maladministration are therefor likely to be minimal. More generally, GCSE grades are known to be important for future educational options (e.g. university entry) and for labour market outcomes (Chowdry et al., 2013). Our outcome measure is therefore of material significance. A potential limitation could be that as these grades reflect young people's knowledge and skills as defined by the English national curricula they might not capture certain skills that inquiry-based methods may have a particular impact upon (e.g. being able to conduct practical experiments independently).

We therefore use PISA science scores as a secondary outcome. Our analysis using this alternative outcome measure is presented in Appendix B and online supplementary material C. The advantage of PISA scores is that they are meant to reflect young people's functional achievement in science including their ability to apply science skills independently in "real-world" situations. They have also been shown to be linked to future educational and labour market outcomes (<blind for review>; Bertschy, Cattaneo, Wolter, 2009). However, they also have important limitations including being based upon a low-stakes test and being measured concurrently with our inquiry teaching variable.

Operationalising inquiry-based teaching

As part of the PISA study, participating children also complete a background questionnaire including questions relating specifically to their science classes. Our interpretation is that

[3] In England, different types of science GCSE examinations are available, and are sat by different pupils. For instance, some schools take triple-science (separate qualifications for Biology, Chemistry and Physics) while others take an integrated science course (counted as two qualifications). We account for such differences in our analysis by using the science pillar score, and by including controls for science course type in our statistical models.

students are most likely to respond in reference to their current (Year 11) science classes. The primary questions of interest in this paper are drawn from the PISA inquiry-based teaching index, all of which are measured on a four-point scale (every lesson, most lessons, some lessons, never or hardly ever):

When learning science topics at school, how often do the following activities occur?

1. Students are given opportunities to explain their ideas.
2. Students spend time in the laboratory doing practical experiments.
3. Students are required to argue about science questions.
4. Students are asked to draw conclusions from an experiment they have conducted.
5. The teacher explains how a science idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties).
6. Students are allowed to design their own experiments.
7. There is a class debate about investigations.
8. The teacher clearly explains the relevance of science concepts to our lives.
9. Students are asked to do an investigation to test ideas.

Recall that we define inquiry teaching using the National Science Education Standards (NSES) definition introduced in the introduction. Table 1 decomposes that definition into its six constituent parts and shows how the nine items from the PISA inquiry-based teaching scale map against them. The table shows that each of the six components of the NSES definition are measured by two or more items from the inquiry-based teaching scale, suggesting that we have achieved good coverage of the inquiry teaching construct. Table 1 also shows that two of the nine items from the PISA scale are not relevant to the NSES definition of inquiry. Item 5, for example, involves teachers explaining the application of science concepts to pupils. This is more akin to teacher-directed instruction, as these methods imply pupils will be receiving knowledge directly from their teachers. A similar objection may also be applied to Item 8, which measures whether teachers explain the relevance of science to pupils. We therefore leave these two items out of our main measure of inquiry teaching, though we do make available online supplementary material D in which our main results are reproduced with all nine items included.

<<Table 1>>

Operationalising inquiry-based teaching

In line with the research reviewed in the introduction, we also explore whether the level of guidance provided moderates the impact of inquiry-based teaching. Recall that guidance can take the form of either constraints placed around the investigation being conducted or targeted prompts/advice delivered at the time a student gets stuck (De Jong & Lazonder, 2014).

Constraints help reduce extraneous cognitive load by cutting out inessential information, thus reducing demands on limited working memory. Cognitive load theorists refer to this as the redundancy effect (Sweller, Ayres, & Kalyuga, 2011c). We measure one important type of constraint using a question drawn from ST103. This question assesses how often a whole class discussion takes place with the teacher (measured on a four-point scale from "Every lesson" to "Never or hardly ever"). By constraining discussion, teachers reduce the number of conjectures, explanations and misconceptions that pupils encounter from other pupils, thus reducing extraneous cognitive load. We do not use other questions from the ST103 scale (e.g. "The teacher discusses our questions") because they do not have any clear theoretical relationship to reducing cognitive load.

Targeted prompts or advice help reduce extraneous cognitive load by helping pupils allocate limited working memory to the most important parts of a task or solution space (Sweller, Ayres, & Kalyuga, 2011d). This can be achieved by teachers modelling a solution, providing cues, or prompting pupils to attend to certain bits of information (Martin, 2016). We measure this using three questions assessing the extent to which teachers provide additional help when students need it, particularly in terms of how to improve their performance. These three questions are drawn from ST100 and ST104 scales. We note that these three items are similar to those on the Load Reduction Instruction Scale, which is specifically designed to measure whether teachers successfully manage cognitive load (Martin & Evans, 2018). We do not use other items from the ST100 or ST104 scales (e.g. "The teacher shows an interest in every student's learning" or "The teacher tells me how I am performing in this course") because they do not have any clear theoretical relationship to reducing cognitive load.

In summary, the four questions we use to measure guidance are:

1. A whole class discussion takes place with the teacher[4] (ST103Q03)
2. The teacher gives students extra help when they need it (ST100Q02)
3. The teacher tells me how to improve my performance (ST104Q04)
4. The teacher advises me how to reach my learning goals (ST104Q05)

We reverse score the first question and then combine these four questions into a single score. We then split the pupils in the data into two groups based on whether they have above or below average guidance scores. For transparency, online supplementary material F provides analogous results for each of the individual items within the ST100, ST103 and ST104 question batteries.

Methods and procedures

The PISA study employs a stratified and clustered sample design. Within England, the country is first divided into different groups (strata) based upon region and school type. Within each of these strata, schools are then randomly sampled with probability to proportional to size. Both the stratification and clustering have implications for the estimation of standard errors. There are different ways that this can be accounted for within analyses of the PISA data. The survey organisers recommend using the Balanced-Repeated-Replication (BRR) weights, which are provided with the data (Avvisati & Keslair 2014; OECD, 2009). These BRR weights are based upon a resampling method (similar to jackknife or bootstrap) and allow the impact of both the stratification and clustering to be incorporated into the estimated standard error[5]. These recommendations were applied throughout our analyses.

We began by constructing our inquiry-based teaching scale using the seven items discussed in the data section above (also see Table 1). First, an item-response theory (IRT) model was estimated including these seven items as measures of the latent inquiry-teaching construct. Expected A Posteriori (EAP) estimates were then created for each student in the dataset capturing the amount on inquiry teaching that they received. These were then standardised to have mean zero and a standard deviation of one across our population of interest. The

[4] This item has been reverse coded.
[5] Alternative methods to account for complex survey designs (e.g. estimation of multi-level models) only capture the impact of clustering and not the impact of the survey stratification per se. See (<*blind for review*>) for further details.

correlation between this scale and the OECD's standard inquiry-teaching scale (IBTEACH), which includes nine items rather than the seven defining our construct, was approximately 0.95.

Using the data described in the previous section, we then investigated the conditional association between our inquiry-based teaching scale and students' GCSE science grades (science pillar points score). We did this by estimating a series of Ordinary Least Squares regression models which included demographic, prior achievement and school-level controls. Imputation was used to account for the small amounts of item non-response. Specifically, models were estimated of the form:

$$GCSE_{ij} = \alpha + \beta.IBTEACH_{ij} + \gamma.D_{ij} + \delta.PISA_{ij} + \tau.KS2_{ij} + \sigma.Course_{ij} + \theta.Scales_{ij} + \mu_j + \varepsilon_i \quad (1)$$

Where:

$GCSE_{ij}$ = Pupils' GCSE science grades. This was measured via the science pillar points score and was standardised to have a mean of zero and standard deviation of 1.

$IBTEACH_{ij}$ = These were a set of dummy variables referring to quartiles of the inquiry-based teaching scale. The bottom quartile (infrequent use of inquiry-based teaching) was set as the reference group. The number of pupils in each quartile was 1,139 (bottom quartile), 1,169 (Q2), 1,159 (Q3) and 1,138 (top quartile)[6].

D_{ij} = A vector of demographic characteristics (e.g. gender, socio-economic status, immigrant group).

$Course_{ij}$ = A vector of controls for the type of science course the student was enrolled in at school (e.g. triple science, double science, etc).

$PISA_{ij}$ = Children's scores on the PISA science test, including the 'content' sub-domains[7].

$KS2_{ij}$ = Key Stage 2 maths point score, reading point score and teacher-assessed science level.

$Scales_{ij}$ = A vector of controls for other factors that potentially impact upon their GCSE science grades but not themselves likely to be influenced by inquiry-based teaching practises.

[6] As the final student weight was applied, the raw number of students within each quartile varies.
[7] Our models controlled for all ten plausible values. This provided the most extensive possible control using the PISA data for each child's ability in science.

μ_j = School-fixed effects.

i = Student i.

j = School j.

ε_i = Error term.

It was also important to consider functional form, as recent research has suggests that the association between inquiry instruction and science achievement may be non-linear (Teig, Scherer, & Nilsen, 2018). There were several different ways non-linear relationships could have been incorporated into the analysis, such as via the inclusion of a quadratic term (as per Teig et al., 2018), using non-parametric regression methods, or by categorising the variable (e.g. dividing the sample into quartiles based upon the IBTEACH scale) and entering these groups as dummy variables into the model. Our main analysis used the latter approach as it allowed us to effectively investigate whether non-linearities were present while also facilitating simple presentation and interpretation of the results. In online supplementary material D, we provide results using alternative methods of accounting for non-linearity. Our overall substantive conclusions are robust to whichever method of accounting for non-linearities is used.

A number of specifications of this model were estimated to illustrate how parameter estimates changed with the addition of extra control variables. Our first specification included basic demographic characteristics (D_{ij}) only. Controls for prior achievement ($PISA_{ij}$ and $KS2_{ij}$) and type of science course studied (e.g. double or triple science) were then added in specification 2 with the school fixed effects (μ_j) added in specification 3. Finally, Model 4 controlled for a further set of potentially confounding characteristics which may have been independently associated with GCSE science performance but are unlikely to have influenced the frequency with which science teachers used an inquiry-based approach. This included (a) the number of minutes timetabled for science in school each week; (b) children's sense of belonging in school; (c) children's anxiety about taking tests; (d) the extent of emotional support children received from their parents; (e) before and after school activities and (f) children's perceptions of whether their science teacher treated them fairly.

Our preferred model is the one which included the full set of controls included in regression model (1). The β parameter from this model captures the association inquiry-based teaching

had with pupils' GCSE grades, given that they had the same demographic background, attended the same school, achieved the same Key Stage 2 scores, performed equally well on the PISA science tests (taken just six months earlier) and had the same sense of belonging in school, anxiety about examinations, did similar before/after school activities and received the same support from their parents. Hence although we cannot claim that this strategy will produce causal estimates, the list of control variables in our preferred specification is extensive. Our results are therefor likely to provide a reasonable indication of whether inquiry-based teaching is independently associated with the academic progress Year 11 pupils make over this critical six-month period. As well as estimating this model based upon the full sample, we also examined potential heterogeneous effects by re-estimating model (1) for specific sub-groups. This included conducting separate analyses by (a) gender, (b) socio-economic status (as measured by the PISA Economic, Social and Cultural Status (ESCS) index) and (c) prior achievement[8].

As well as having conducted the analysis using the final PISA inquiry-based teaching scale, in the supplementary material we also provide a breakdown of estimates using each individual question. Take the question '*students are allowed to design their own experiments'*. We re-estimated equation (1) removing the inquiry teaching quartiles and included students' responses to this question in its place. This was done for each of item belonging to the inquiry-based teaching scale (listed above). These estimates reveal whether any specific type of inquiry-based classroom activity is particularly strongly associated with science attainment growth. We also investigated whether our measures of guidance moderated the relationship between inquiry and pupil attainment.

Finally, we attempted to address two important limitations with the inquiry-based teaching measure. The first is that it only provides information about how *frequently* students said inquiry-based teaching was used, but nothing about the *quality* of how it was implemented. Although we used nationally representative data, meaning our estimates should have captured the average effect of inquiry-based teaching as it was actually implemented in England's schools, conducting further robustness tests around this issue was important. We therefore drew upon the PISA 'disciplinary climate' scale. This captures students' reports of how much

[8] The ESCS index was created by the OECD to provide an overall measure of young people's family background. It incorporates information on parental education, occupation and household possessions. In our analysis we divided pupils according to this index into three socio-economic status groups: low, average and high.

disorder there was in their science classrooms, in response to statements such as '*students cannot work well'*, '*there is noise and disorder'* and '*students don't listen to what the teacher says'*. Our reasoning was that in classrooms where discipline was a problem, it is unlikely that inquiry-based teaching practises were being implemented well. We therefore explored whether the effect of inquiry-based teaching differed according to the disciplinary climate of the science classroom. We conducted this analysis using measures of disciplinary climate at the student-level (i.e. using students' own reports) or at the school-level (i.e. using the school-average of the disciplinary climate scale).

The second limitation of our inquiry-based teaching variable is that it may be subject to some measurement error. Consequently, Appendix A provides alternative estimates where we have combined information across all sampled pupils within each school to create an alternative measure of inquiry-based instruction[9]. Specifically, we took the school mean of the IBTEACH variable, under the assumption that any under or over reporting of inquiry-based teaching by pupils would likely cancel out within a given school. To trail the key finding from Appendix A, when we used this alternative measure of inquiry-based teaching there was almost no change to the results or the substantive conclusions that we reached.

All analyses were conducted using Stata version 14 (StataCorp 2015) with the statistical code available from <*blind for review*>.

Results

The overall effect of inquiry-based teaching

Table 2 presents results across our four model specifications. As the outcome variable has been standardised, all estimates can be interpreted in terms of an effect size. In the baseline specification (M1), including demographic characteristics only, we find a moderate positive association between being in the second and third inquiry-based teaching quartile compared to the bottom quartile as the reference group. However, there is no difference in GCSE science grades between students who receive little inquiry-based teaching (bottom quartile) versus those who receive a lot (top quartile).

Of course, a key issue with model M1 is that it includes only a limited set of controls for potentially confounding characteristics. Specification M2 therefore adds controls for prior

[9] These alternative estimates exclude the school fixed-effects, because these are collinear with the collapsed school-level inquiry-based teaching variable that we derived.

achievement, as measured by Key Stage 2 grades and PISA science scores, along with information on the type of science courses children are studying (e.g. double versus triple science). The estimates indicate a small positive effect of inquiry-based instruction upon the progress young people make in science during Year 11. Emphasis should however be placed upon the word small. A substantial difference in the frequency of inquiry-based instruction students receive (top versus bottom quartile) is associated with only a 0.1 standard deviation increase in GCSE science scores. This is the equivalent of an increase of approximately one-tenth of a single GCSE science grade.

A similar result holds when school fixed-effects are added in model M3. The coefficient for the top quartile versus the bottom quartile remains stable in magnitude and statistical significance. Moreover, once additional controls for other characteristics of the pupils have been added in model M4 (e.g. their sense of belonging in school, test anxiety, parental support) none of the inquiry-based teaching quartiles remain statistically significant. Together, Table 2 therefore suggests that frequency of inquiry-based teaching practises has little overall association with Year 11 pupils' academic achievement.

In the online supplementary material we have also investigated whether there are some specific inquiry practises which are positively associated with students' outcomes. Specifically, we have investigated the effect of undertaking one of the seven inquiry practises in 'some', 'most' and 'every' lesson relative to 'never/hardly ever' as the reference group. Consistent with the findings presented throughout this section, the vast majority of effect size estimates are small (below 0.1) and statistically insignificant. This further supports our conclusion that Year 11 science teachers who regularly use inquiry-based instruction methods during their lessons are unlikely to boost their students' GCSE grades.

The effect of guided practices

Although we find no consistent effects across the inquiry-based teaching scale, it may be that certain types of inquiry-based teaching improve attainment. This includes guided-inquiry practices. Unfortunately, the PISA background questionnaire does not provide information on whether the inquiry teaching practises of science teachers were guided or not. Respondents were however asked about the extent to which their science teacher provided guidance as part of their teaching practice in general. Of course, it is possible that such guidance may not be related to their inquiry activities. Nevertheless, we believe further investigation of guidance as a moderator remains worthwhile, despite this limitation.

Table 3 shows the coefficient from our preferred specification (Model 4) across the inquiry quartiles, split by the whether a pupil reports these four types of guidance occur with a high or low frequency[10]. The table shows no relationship between any quartile of inquiry-based teaching for the pupils reporting low levels of guidance. Among the pupils reporting high guidance however, the third and fourth quartile of inquiry is occasionally associated with increased attainment with effect sizes ranging between 0.02 and 0.16. Interestingly, the coefficients are slightly larger in the third quartile than in the fourth. In summary, neither high inquiry with low guidance, or high guidance with low inquiry are related to improved science attainment. There is however, some tentative evidence that high inquiry delivered in conjunction with high guidance may have a small positive impact upon science achievement. This pattern of findings is consistent with the predictions of cognitive load theory.

Estimates for sub-groups

Although there is no effect of inquiry-based teaching on average, this could mask differential effects across sub-groups. For instance, there might be a benefit for high-achieving pupils who are able to explore their ideas more during their science lessons, but a negative effect for lower-achievers who still do not have a firm grasp of the important concepts.

Table 4 provides no evidence that this is the case. Based upon model specification 4, it illustrates whether the 'impact' of inquiry-based teaching varies by gender, socio-economic status and prior achievement (as measured by PISA scores). For all groups, the estimated effects are small and fail to reach statistical significance at the five percent level. There is hence no evidence that inquiry-based teaching methods are particularly effective for any of these sub-groups.

Implementation issues?

One potential explanation for our null results is that science teachers in England are (on average) failing to deliver inquiry-based science teaching methods appropriately. Hence we may find a greater impact if we could also account for the *quality* of implementation. Although a direct measure of quality is not available, we do have access to pupils' perceptions of the disciplinary climate within their science classroom. We argue that, given the practical nature

[10] Online supplementary material F illustrates how results differ across each item included within the ST100, S103 and ST104 battery of questions. Each of these batteries asks pupils to provide responses about the teaching approaches used by their science teacher.

of inquiry-based approaches, poor discipline in the classroom (e.g. lots of noise and disorder, pupils not listening to their teacher when performing experiments) is likely to be a good marker of whether implementation of such teaching methods is reasonably good or rather poor.

Table 5 therefore presents results separately for pupils where the disciplinary climate in the science classroom is good, average or poor based upon thirds of the PISA 'disciplinary climate' scale. The top panel refers to when we divide pupils into different groups based upon their own reports of the disciplinary climate in their science classroom. The bottom panel, on the other hand, presents results where entire schools have been divided into good, average and poor disciplinary groups[11]. Once again, all estimates continue to be small and statistically insignificant. Even in classrooms with a good disciplinary climate, the difference in science achievement between pupils in the top and bottom inquiry teaching quartile is just an effect size of 0.1.

Further robustness tests

Appendix B contains some additional robustness tests. One potential explanation for our finding of null effects is that our preferred value-added model specification captures the amount of progress Year 11 pupils make over a relatively short time horizon. We therefore run an additional version of our model in which we use only Key Stage 2 scores (not PISA scores) as the prior achievement measure in our statistical models. Hence our estimates now illustrate how inquiry-based teaching methods are associated with the progress pupils make over a five-year time horizon between the end of primary school (when they take Key Stage 2 tests) and the end of secondary school (when they sit GCSE examinations). In our preferred specification (Model 4) the coefficient on being in the second or third quartile is now statistically significant, but the effect size remains below 0.1. The fourth quartile has a slightly smaller coefficient than the third and is not statistically significant at conventional levels. This is a small effect size given we are now considering the impact of such methods over a five-year time horizon.

Another potential explanation for our results is that inquiry-based teaching methods may have more impact upon 'functional' real-world science skills – i.e. those skills that PISA attempts to measure – rather than performance on a curriculum-based test of scientific knowledge such as GCSEs. We therefore switch to using PISA science scores instead of using GCSE science grades as our outcome measure (again, using Key Stage 2 as our only prior achievement

[11] This is based upon the school average of the disciplinary climate scale.

measures). In our preferred specification (Model 4), only the top quartile is statistically significant – but the effect is negative (0.10 standard deviations *lower* than for students in the bottom inquiry teaching quartile). Consequently, our substantive conclusions are also robust to the measure of science achievement used.

Discussion

Science teachers face important decisions about how to design instruction for their pupils. One prominent school of thought, inquiry learning, holds that students learn science best by conducting experiments to answer research questions. Thus, teachers should design opportunities for students to acquire knowledge through investigations, rather than providing it to them directly. This research set out to provide new evidence on the effectiveness of inquiry-based teaching, specific components of inquiry-based teaching and inquiry-based teaching coupled with more or less guidance. The results indicate that inquiry-based teaching has a very weak relationship with attainment in science – and that any positive effects are confined to moderate levels of inquiry combined with high levels of guidance. High levels of inquiry or unguided inquiry have no relationship with attainment at all. These results are consistent with existing literature, which tends to find that inquiry is less effective than more direct forms of instruction (Kirschner et al., 2006; Stockard et al., 2018; Alfieri et al., 2011) except for in cases where the inquiry is highly guided (Hmelo-Silver et al., 2007; Lazonder & Harmsen, 2016).

A reasonable objection to many evaluations of inquiry-based teaching is that the outcome measures used fail to capture the functional or real-world science skills which are, in part, what inquiry-teaching aims to inculcate in students. However, our findings seem to rule out this interpretation because we find small or zero effects of inquiry both when using traditional, high-stakes GCSE examination and when using PISA test scores (which are designed to measure such real-world skills). Another possible explanation is that our measure of inquiry-based teaching focuses upon the frequency with which inquiry practices are used, rather than the quality. We attempted to test this by checking whether classroom discipline moderated the relationship between inquiry and attainment. While this is clearly a very indirect measure of the quality of inquiry teaching, we found no evidence that discipline moderated the

relationship. In any case, our data comes from a large representative sample of teachers and pupils meaning we are evaluating the quality of inquiry teaching as currently displayed in schools in England. Our measures of inquiry-based teaching therefore have ecological validity. Our interpretation of these results is therefore that the benefit of allowing pupils to acquire their own knowledge through investigation are small and, consistent with the theory reviewed in the introduction, can easily be cancelled out by the additional cognitive load involved in conducting such investigations.

This research has implications for the practice of science teaching. In particular, it suggests that science teachers should not overuse inquiry methods. The limits apply to both of the dimensions of inquiry identified by Rönnebeck et al. (2016) – inquiry activities should be used in moderation and accompanied by high levels of guidance. Teachers can achieve the latter by reducing the scope or number of decisions involved in investigations conducted by pupils, modelling solutions or worked examples, or providing carefully timed prompts and heuristics to direct pupils' attention towards relevant aspects of the tasks (Rosenshine, 2012; De Jong & Lazonder, 2014).

Limitation and future research

These findings should, of course, be considered in light of the limitations of this study. First, our focus has been upon the frequency which inquiry-based methods are taught and not about the quality by which they are delivered. While we have conducted some important robustness tests surrounding this issue, we cannot rule out the possibility that inquiry-based approaches may be able to improve young people's achievement when delivered unusually well. Likewise, we also do not know the quality of the non-inquiry methods against which we are making our comparisons. Second, our measure of inquiry is based upon student reports and could therefore be subject to some reporting and recall inaccuracies. There is no particular reason to believe that young people would struggle to report such information and the reliability of the scale reported in the technical documentation is relatively high[12]. Having said that, we cannot rule out measurement error having some impact upon our results. Third, PISA only collects information about teacher-guidance in general and not specifically about the use of guided-inquiry approaches. Although we have provided some insight into this issue (under the

[12] Across the UK, the reliability (Cronbach's alpha) of the inquiry-based teaching scale is 0.86 (see OECD 2017: Table 16.29).

assumption that teachers who use guidance more in general also provide more guidance within inquiry activities) further work using more precise measures is needed. Finally, as with all observational studies, our estimates refer to conditional observations only and do not necessarily capture cause and effect. Although our longitudinal analysis has conditioned upon a wide array of potential confounding factors – including measures of prior achievement not available within the international PISA database – the presence of potentially important unobserved factors cannot be ruled out.

Ideally, future research will attempt to measure the quality of the inquiry-based teaching rather than just the quantity. This might include the development and validation of observation rubrics. In addition, rather than considering inquiry instruction in isolation, its impact should be investigated in conjunction with the wide range of other approaches that teachers use. Ideally, the mix of inquiry and other methods employed by teachers would be modelled in order to get a cleaner estimate of the effect of inquiry instruction upon pupil achievement. Although this is beyond the scope of this paper, and the data we currently have available, future research – possibly using structural equation modelling – should explore this possibility. Additional outcome data should also be collected where possible. For example, although we are able to employ both traditional measures of pupil attainment based upon national curricula, as well as the application of science to real world problems assessed in PISA, it would also be of interest to investigate other outcomes. For example, procedural knowledge could be assessed through controlled assessments in which pupils are observed conducting practical work. Investigating the effect of inquiry methods on motivation is also important. For example, even some cognitive load theorists (Martin, 2016) acknowledge that providing students with greater autonomy in the classroom should improve their motivation. It is therefore important to understand how inquiry-teaching is related to pupils' interest and engagement in science, as well as their decision as to whether they continue studying science beyond compulsory education.

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1-18.

Association of Science Education [ASE] (2009) Getting Practical. Accessed 21/11/18: https://www.ase.org.uk/professional-development/getting-practical/

Association of Science Education [ASE] (2018). Scientific enquiry in the UK. Accessed 21/11/18: https://www.ase.org.uk/resources/scientific-enquiry-in-uk

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

Avvisati, F., & Keslair, F. (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*. Statistical Software Components S457918, Boston College Department of Economics.

Bell, R. L., Smetana, L., & Binns, I. (2005). Simplifying inquiry instruction. *The Science Teacher, 72*(7), 30-33.

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349-377.

Bertschy, K., Cattaneo, M, & Wolter, S. (2009). PISA and the transition into the labour market. *Labour, 23*, 111-137.

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.

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2013). Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society Series A, 176*(2), 431-457.

De Jong, T., & Lazonder, A. W. (2014). The guided discovery learning principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 371–390). New York, NY: Cambridge University Press.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300-329.

Gibb, N. (2017). Nick Gibb: the evidence in favour of teacher-led instruction. Accessed 06/08/2018: https://www.gov.uk/government/speeches/nick-gibb-the-evidence-in-favour-of-teacher-led-instruction

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*(4), 267-321.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clarke (2006). *Educational Psychologist*, *42*(2), 99-107.

Hodson, D. (2014). Learning science, learning about science, doing science: Different goals demand different learning methods. *International Journal of Science Education*, *36*(15), 2534-2553.

Holman, J. (2017). *Good practical science*. London: Gatsby.

Kirschner, P. A. (1992). Epistemology, practical work and academic skills in science education. *Science & Education*, *1*(3), 273-299.

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. (2013). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences, 110*(3), 14075-14080.
Lazonder, A., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning. Effects of guidance. *Review of Educational Research, 86*(3), 681-718.
Martin, A. J. (2016). *Using Load Reduction Instruction (LRI) to boost motivation and engagement*. Leicester: British Psychological Society.
Martin, A. J., & Evans, P. (2018). Load reduction instruction: Exploring a framework that assesses explicit instruction through to independent learning. *Teaching and Teacher Education*, *73*, 203-214.
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.
National Research Council [NRC] (1996). *National science education standards*. Washington, DC: The National Academies Press.
National Research Council [NRC] (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: The National Academies Press.
National Research Council [NRC] (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
OECD (2006). *Contextual framework for PISA 2006*. Paris: OECD Publishing. Accessed 21/11/2018 from: https://www.acer.org/files/pisa2006_context_framework.pdf
OECD (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. Paris: OECD Publishing.
OECD (2010). *The High Cost of Low Educational Performance*. Paris: OECD Publishing.
OECD (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: OECD Publishing.
OECD (2017). *PISA 2015 technical report*. Paris: OECD Publishing.
Ofsted (2013). *Maintaining Curiosity: a survey into science education in schools*. London: Ofsted.
Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., ... & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47-61.
Pressley, M., Roehrig, A.D., Raphael, L., Dolezal, S., Bohn, C., Mohan, L., Wharton-McDonald, R., Bogner, K. & Hogan, K. (2003). Teaching processes in elementary and secondary education. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of educational psychology* (pp. 153-176). New Jersey: John Wiley & Sons.
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*. Brussels: European Commission.
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.
Rosenshine, B. (2012). Principles of instruction: Research-based strategies that all teachers should know. *American Educator, 36*(1), 12-19.
Rutherford, F. J. (1964). The role of inquiry in science teaching. *Journal of Research in Science Teaching, 2*(2)*,* 80–84.

Rutherford, E. J., & Ahlgren, A. (1991). The Need for Scientific Literacy. *Thinking: The Journal of Philosophy for Children*, *9*(4), 13-19.

Schwab, J. J. (1962). The teaching of science as inquiry. In J. J. Schwab, & P. F. Brandwein (Eds.), *The teaching of science* (pp. 3–103). Cambridge, MA: Harvard University Press.

Sjøberg, S. (2015). PISA and global educational governance – a critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(1), 111-127.

StataCorp (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507.

Sweller, J., Ayres, P., & Kalyuga, S. (2011a). *Cognitive load theory*. Berlin: Springer.

Sweller, J., Ayres, P., & Kalyuga, S. (2011b). Intrinsic and extraneous cognitive load. In *Cognitive load theory* (pp. 57-69). Springer, New York, NY.

Sweller, J., Ayres, P., & Kalyuga, S. (2011c). The redundancy effect. In *Cognitive load theory* (pp. 141-154). Springer, New York, NY.

Sweller, J., Ayres, P., & Kalyuga, S. (2011d). Interacting with the external environment: The narrow limits of change principle and the environmental organising and linking principle. In *Cognitive Load Theory* (pp. 39-53). Springer, New York, NY.

Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction*, *56*, 20-29.

Varsakelis, N. C. (2006). Education, political institutions and innovative activity: A cross-country empirical investigation. *Research Policy*, *35*(7), 1083-1090.

Zhang, L. (2016). Is inquiry-based science teaching worth the effort? *Science & Education, 25*(7), 897-915.

Zhang, L. (2018). “Hands-on” plus “inquiry”? Effects of withholding answers coupled with physical manipulations on students' learning of energy-related science concepts. *Learning and Instruction*. DOI: https://doi.org/10.1016/j.learninstruc.2018.01.001

Table 1. Operationalising inquiry-based teaching using the NSES definition and items from the PISA inquiry-based teaching scale

	Asking questions	**Planning & conducting investigations**	**Using appropriate techniques to gather data**	**Think about relationship between evidence & explanations**	**Constructing & analysing alternative explanations**	**Communicating scientific arguments**
Item 1: Students explain ideas						✓
Item 2: Students do lab experiments		✓	✓			
Item 3: Argue about science questions	✓				✓	✓
Item 4: Draw conclusions from experiments		✓	✓	✓		
Item 5: Teacher explains applications of science						
Item 6: Students design experiments	✓	✓				
Item 7: Class debate investigations				✓	✓	✓
Item 8: Teacher explains relevance of science						
Item 9: Students test ideas through investigation		✓	✓	✓		

Table 2. Estimated association between inquiry-based teaching and students' GCSE science grades

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale								
Bottom quartile (Reference)	Reference		Reference		Reference		Reference	
Second quartile	**0.11***	**0.04**	**0.06***	**0.03**	**0.05***	**0.02**	0.02	0.02
Third quartile	**0.16***	**0.04**	**0.09***	**0.03**	**0.06***	**0.03**	0.02	0.03
Top quartile (extensive use)	-0.05	0.04	**0.10***	**0.03**	**0.10***	**0.03**	0.06	0.03
Observations	4,361		4,361		4,361		4,361	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
PISA science scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
School fixed effects	-		-		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

Notes: All figures in the effect column can be interpreted in terms of effect sizes. SE = Standard error. Bold coefficients with a * indicate $p<0.05$.

Table 3. Estimated association between different types of inquiry-based teaching practices and students GCSE science grades

Guidance Measures	Low Guidance			High guidance		
	Q2	Q3	Q4	Q2	Q3	Q4
The teacher gives students extra help when they need it	**0.17***	0.09	-0.10	0.02	0.04	0.02
	(0.07)	(0.10)	(0.12)	(0.04)	(0.04)	(0.04)
A whole class discussion takes place with the teacher	0.02	0.01	-0.02	0.09	**0.16***	0.12
	(0.04)	(0.04)	(0.05)	(0.07)	(0.07)	(0.06)
The teacher tells me how to improve my performance	0.02	0.03	0.03	0.07	0.09	0.06
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
The teacher advises me how to reach my learning goals	0.03	0.01	0.02	0.09	**0.12***	0.08
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)

Notes: All coefficients can be interpreted in terms of effect sizes, with the lowest discovery quartile as the reference group. Standard errors are shown in parentheses. Bold coefficients with a * indicate $p<0.05$. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included). Q2, Q3 and Q4 refers to quartiles of the inquiry teaching scale.

Table 4. The estimated impact of inquiry-based teaching practices for different sub-groups

		Second quartile		Third quartile		Top quartile	
	N	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Gender							
Girls	2,125	0.01	0.03	0.03	0.04	0.03	0.05
Boys	2,236	0.00	0.04	-0.02	0.04	0.05	0.04
Socio-economic status							
Low SES	1,279	**0.13***	**0.05**	**0.15***	**0.05**	0.09	0.06
Average SES	1,404	-0.06	0.05	-0.09	0.05	0.00	0.05
High SES	1,491	0.02	0.05	0.00	0.05	**0.11***	**0.05**
Science achievement							
Low-achieving	1,216	0.03	0.05	-0.04	0.06	-0.03	0.06
Average-achieving	1,473	-0.01	0.04	0.02	0.04	0.06	0.04
High-achieving	1,672	0.02	0.04	0.06	0.05	0.09	0.06

Notes: All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. Bold coefficients with a * indicate $p<0.05$. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included). Science-achievement groups based upon top third, middle third and bottom third of pupils in England on the PISA science scale (using the first plausible value). Socio-economic status (SES) based upon thirds of the ESCS index (where data available).

Table 5. The estimated impact of inquiry-based teaching practices for schools with different disciplinary climates

		Second quartile		Third quartile		Top quartile	
	N	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Science class discipline (pupil report)							
Poor discipline	1,373	0.03	0.05	0.04	0.06	0.05	0.06
Average discipline	1,366	-0.01	0.04	0.00	0.05	0.02	0.05
Good-discipline	1,358	-0.04	0.06	0.02	0.04	0.07	0.06
Science class discipline (school-average report)							
Poor discipline	1,332	0.04	0.05	0.06	0.05	0.03	0.06
Average discipline	1,396	-0.02	0.05	-0.04	0.05	0.05	0.05
Good-discipline	1,369	0.07	0.05	**0.10***	**0.04**	**0.12***	**0.05**

Notes: All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. Bold coefficients with a * indicate $p<0.05$. Students/schools have been divided into three groups, based upon students' reports of the disciplinary climate within their science classes. This is based upon the PISA science 'discipline' scale. Top panel refers to results where students have been divided into thirds based upon their own reports. Bottom panel is where we have used the school average of the discipline scale to divide pupils into groups. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included).

Appendix A. Alternative estimates based upon school-average values of the inquiry teaching scale

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale	-0.01	0.03	0.01	0.02	0.01	0.02	0.00	0.02
Observations	**4,318**		**4,318**		**4,318**		**4,318**	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
PISA science scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

Notes: The inquiry-based teaching scale is now based upon the average within the school. It has been entered into the model as a continuous term. Hence estimates refer to standard deviation increases in GCSE science scores per standard deviation increase in the (school-level) inquiry-based teaching scale. All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. Bold coefficients with a * indicate $p<0.05$.

Appendix B. Alternative estimates controlling for Key Stage 2 scores as the only prior achievement variables

(a) GCSE science grades as outcome

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale								
Bottom quartile (Reference)	Reference		Reference		Reference		Reference	
Second quartile	**0.11***	**0.04**	**0.09***	**0.03**	**0.08***	**0.03**	0.04	0.03
Third quartile	**0.16***	**0.04**	**0.12***	**0.03**	**0.09***	**0.03**	0.03	0.03
Top quartile (extensive use)	-0.05	0.04	0.05	0.03	0.05	0.03	0.01	0.03
Observations	4,361		4,361		4,361		4,361	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
School fixed effects	-		-		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

(b) PISA science scores as outcome

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale								
Bottom quartile (Reference)	Reference		Reference		Reference		Reference	
Second quartile	**0.10***	**0.04**	**0.08***	**0.03**	**0.07***	**0.03**	0.04	0.03
Third quartile	**0.08***	**0.04**	0.05	0.03	0.05	0.03	0.02	0.03
Top quartile (extensive use)	**-0.22***	**0.04**	**-0.10***	**0.03**	**-0.09***	**0.03**	**-0.10***	**0.03**
Observations	4,977		4,977		4,977		4,977	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
School fixed effects	-		-		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

Note: All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. Bold coefficients with a * indicate $p<0.05$.

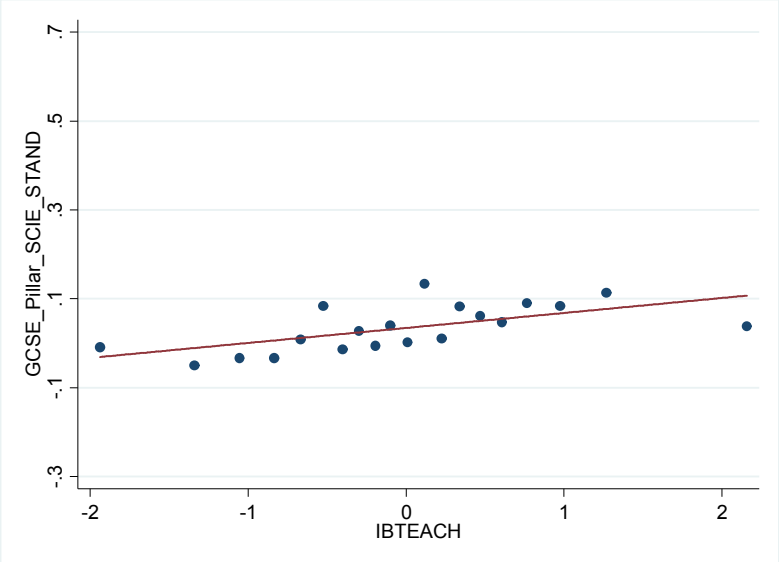
Online supplementary material C. Investigating functional form

In the main body of the paper, we divided the inquiry instruction scale into quartiles and entered this into our OLS regression models as a series of dummy variables. This has the advantage of being straightforward to interpret while also allowing for potential non-linearities in the relationship between inquiry-based instruction and young people's science achievement. However, it also has limitations; the use of quartiles may not fully capture the correct functional form of the relationship while statistical power is also slightly reduced (relative to using a single continuous linear term). In this supplementary material, we consider the issue of functional form in further detail and perform sensitivity analysis for a selection of our results.

GCSE grades

We begin our investigations focusing upon the relationship between the inquiry teaching scale and young people's GCSE science grades. Figure C1 presents a binned scatter plot. This is based upon regression model specification 2, which includes controls for demographic characteristics and prior achievement. The horizontal axis provides the residualised values of the inquiry-based teaching index, while the vertical axis presents the residualised values of young people's science GCSE grades. The red sloping line then illustrates the OLS regression estimate when the inquiry teaching scale is entered as a simple linear term. The blue dots provide a summary of the data points; if these trace the OLS line closely it suggests a simple linear functional form would be appropriate. Alternatively, if the data points do not trace the red OLS regression line, then a more complex functional form may be needed.

Figure C1. Binned scatterplot of IBTEACH vs GCSE science grades (model 2 controls)



In Figure C1, the datapoints clearly follow a linear trajectory and closely follow the fitted OLS regression line. Consequently, functional form for our main analysis does not seem to be a particular issue; even including the inquiry-based teaching scale as a linear term would fit the data well. Moreover, Figure C1 further supports the conclusion we reached in the main body of the paper; the gradient of the fitted regression line is shallow, with little evidence that frequency of inquiry-based teaching is linked to higher levels of science achievement.

These results are formalized in Table C1. The left-hand column illustrate the association between a one standard deviation increase in inquiry-based teaching and young people's science GCSE grades in terms of an effect size. (Again, we estimate our second model which controls for demographic characteristics and prior achievement, but *not* school fixed effects and other characteristics). The effect is clearly small, standing at just 0.035 standard deviations. More importantly, the middle column illustrates how there is almost no change to this result when a quadratic term is added to the model (i.e. we allow there to be a 'curvilinear' relationship). Infact, the quadratic term is extremely small and not statistically significant. This is consistent with the results presented in Figure C1 above.

Table C1. Results using different functional forms. GCSE grades.

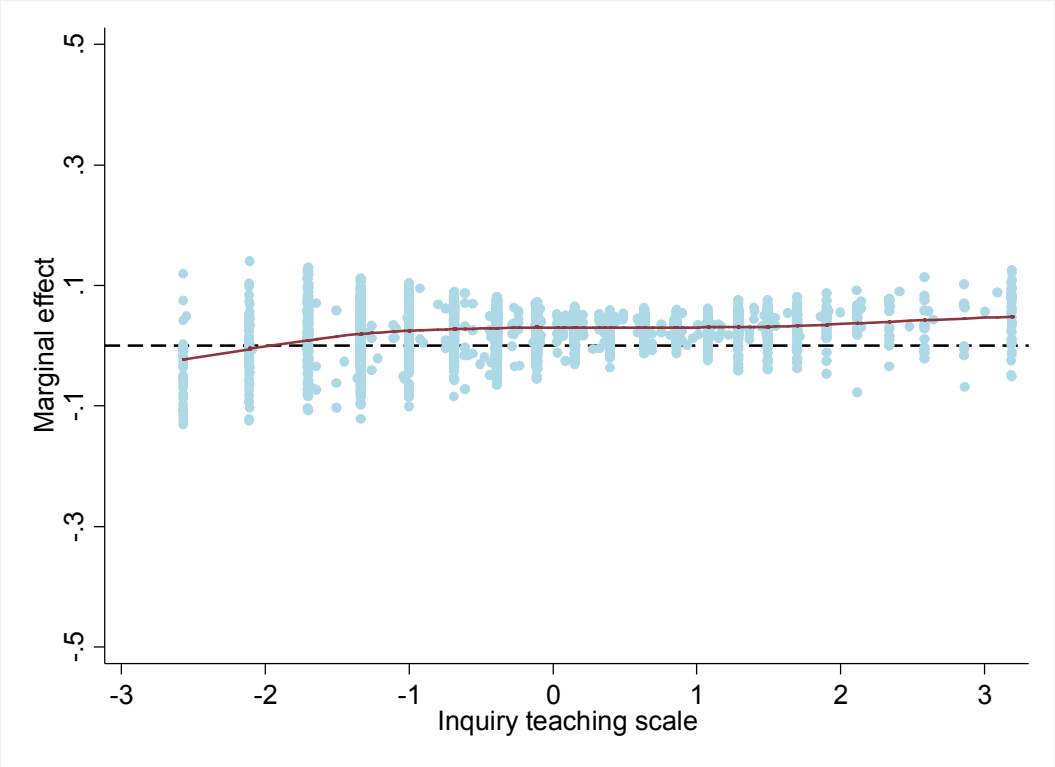
	Linear		**Quadratic**		**Non-parametric**	
	Effect	**SE**	**Effect**	**SE**	**Average marginal Effect**	**SE**
Inquiry-teaching scale						
Main effect	**0.03***	**0.01**	**0.04***	**0.01**	**0.03***	**0.01**
Quadratic effect	N/A	N/A	-0.01	0.01	N/A	N/A

Notes: Model 2 controls. Standard errors not clustered for the average marginal effect.

As a final check, Figure C2 presents results from a non-parametric regression model. The horizontal axis again contains values of the inquiry teaching scale. Meanwhile, the vertical axis presents the "marginal effect" of inquiry-based teaching upon children's science achievement. Positive values suggest that, at that given level of inquiry-based teaching, there is a positive association with young people's achievement. Negative values, on the other hand, suggest that inquiry teaching is associated with a reduction in science achievement. A reference line has been added at zero; the point where inquiry teaching has neither a positive or negative effect.

The main message of Figure C2 is that frequency of inquiry teaching has little baring upon young people's GCSE science grades. The fitted non-parametric regression line is always around zero, without a particularly pronounced positive or negative effect at any point along with inquiry-based teaching scale. Consistent with the results presented thus far, Table C1 also illustrates how the average marginal effect is very small (0.03), strengthening the evidence that frequency of inquiry-based teaching has little impact upon children's performance in their GCSE science exams.

Figure C2. Non-parametric regression results for GCSE science grades



Notes: M2 controls. Figures on y-axis refer to the marginal effect in inquiry teaching in terms of an effect size. Dashed horizontal line plotted where the marginal effect is zero. Solid red line illustrates the non-parametric regression line. Weights not applied.

PISA scores

We now provide analogous results for PISA science scores. Figure C3 begins by presenting estimates from the binned scatterplot. Interestingly, and in contrast to our findings for GCSE science grades, a simple linear term for the IBTEACH scale does not seem appropriate. In particular, note how the data points do not closely resemble the trajectory of the OLS regression line. They, in fact, suggest that that a quadratic (curvilinear) relationship may be more appropriate.

This result is formalised in Table C2. Specifically, in the middle column a quadratic (curvilinear) term is included in the model, which is sizable and statistically significant. This further reiterates the case that choice over functional form is likely to be important when one considers the relationship between inquiry-based teaching and PISA science scores (which was not the case for GCSE grades).

Figure C3. Binned scatterplot of IBTEACH against PISA science scores

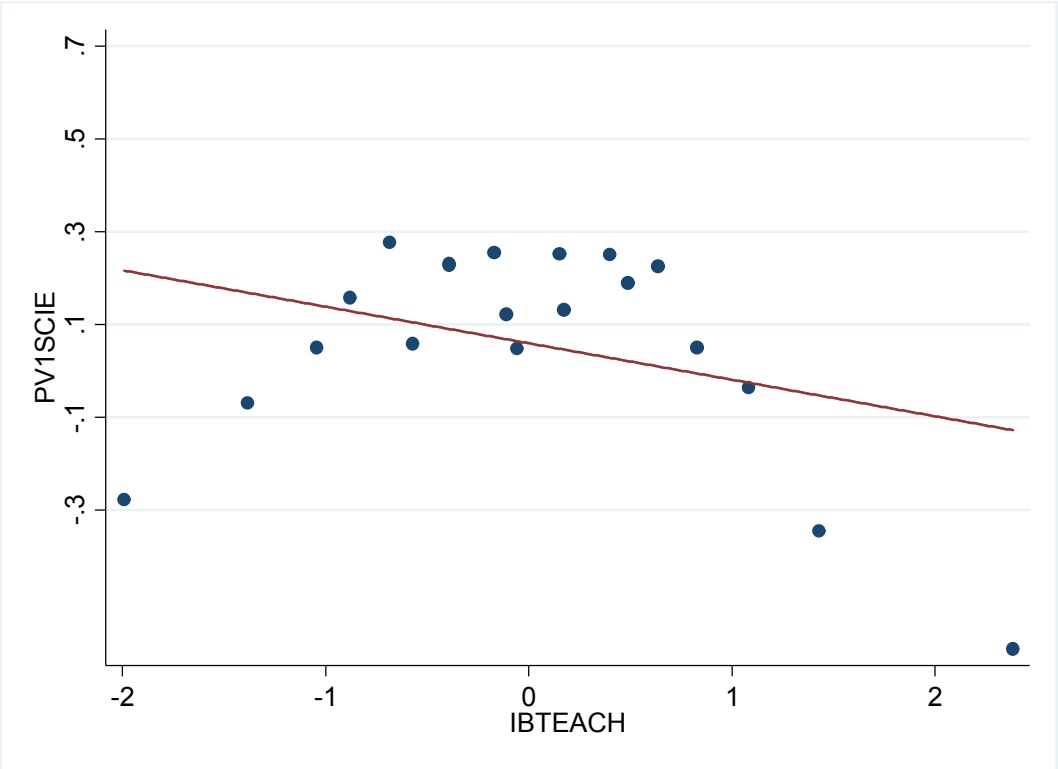


Table C2. Results using different functional forms. PISA scores.

	Linear		Quadratic		Non-parametric	
	Effect	**SE**	**Effect**	**SE**	**Average marginal Effect**	**SE**
Inquiry-teaching scale						
Main effect	**-0.04***	**0.01**	-0.02	0.01	0.00	0.01
Quadratic effect	N/A	N/A	**-0.08***	**0.01**	N/A	N/A

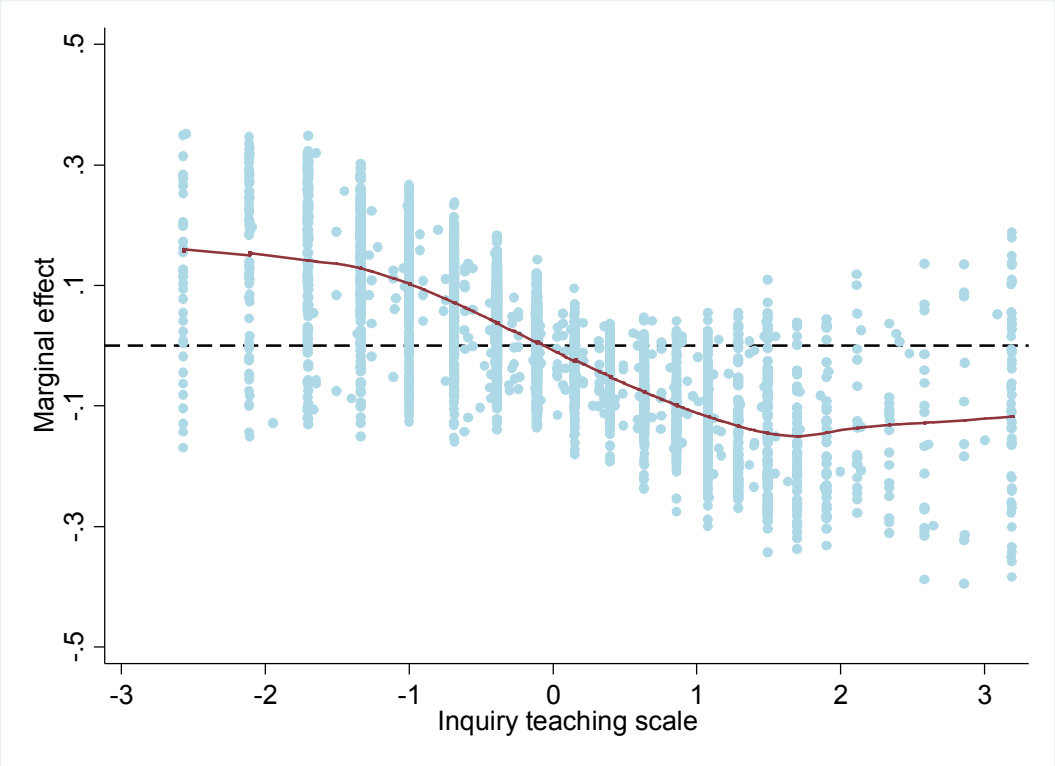
Notes: Model 2 controls.

To understand this relationship in more detail, Figure C4 presents results from a non-parametric regression model. Once again, values of the inquiry-based teaching index run along the horizontal axis, with the "marginal effect" of this teaching approach (reported in terms of an effect size) plotted on the y-axis. A reference line is again included at zero to highlight where the effect of inquiry-teaching turns from positive to negative.

Figure C4 suggests that there is small positive benefit from doing some inquiry instruction versus doing none. This is illustrated by the fact that the fitted non-parametric regression line is greater than zero towards the left-hand side of the graph. However, it should be noted how the effect size remains modest – only around 0.1 standard deviations. In contrast, towards the right-hand side of the graph the marginal effect is weakly negative, with moving from doing a

moderate amount of inquiry-teaching to a lot associated with a 0.1 standard deviation decrease in PISA science scores. As Table C2 illustrates, these positive and negative effects on average cancel one another out, with an average marginal effect of just 0.1.

Figure C4. Non-parametric regression results for PISA science scores



Notes: M2 controls. Figures on y-axis refer to the marginal effect in inquiry teaching in terms of an effect size. Dashed horizontal line plotted where the marginal effect is zero. Solid red line illustrates the non-parametric regression line. Weights not applied.

In summary, while the issue of functional form is of little importance for our analysis of GCSE science grades, it is a more prominent issue with respect to the link between inquiry instruction and PISA science scores. Overall, our analysis suggests that the relationship between inquiry-based teaching and young people's science skills is weak at best. Based upon our analysis, we suggest that there may by some small benefits from teachers doing some inquiry instruction during lessons rather than doing none at all, although even this result is dependent upon the outcome measure used (it holds only for PISA science scores and not GCSE grades).

Online supplementary material D. Alternative estimates using the original nine-item inquiry-based teaching scale (developed by the OECD)

Table D2. Estimated association between inquiry-based teaching and students GCSE science grades

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale								
Bottom quartile (Reference)	Reference		Reference		Reference		Reference	
Second quartile	**0.13***	**0.05**	0.05	0.03	0.03	0.03	0.00	0.03
Third quartile	**0.22***	**0.04**	**0.08***	**0.02**	**0.05***	**0.02**	0.01	0.03
Top quartile (extensive use)	-0.01	0.05	**0.09***	**0.03**	**0.07***	**0.03**	0.03	0.03
Observations	**4,361**		**4,361**		**4,361**		**4,361**	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
PISA science scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
School fixed effects	-		-		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

Notes: All figures in the effect column can be interpreted in terms of effect sizes. SE = Standard error. * indicates statistical significance at the five percent level.

Table D3. Estimated association between different types of inquiry-based teaching practices and students GCSE science grades

	Some			Most			Every		
	%	**Effect**	**SE**	**%**	**Effect**	**SE**	**%**	**Effect**	**SE**
Items									
1. Explain ideas	20%	-0.03	0.05	38%	-0.03	0.04	37%	0.00	0.05
2. Practical experiments	62%	0.04	0.03	**15%**	**0.09***	**0.04**	4%	0.06	0.06
3. Argue about science questions	37%	0.02	0.02	12%	-0.04	0.03	5%	0.05	0.05
4. Conclusions from experiments	45%	0.05	0.04	36%	0.07	0.04	12%	0.05	0.04
5. Apply to phenomena	30%	0.01	0.03	39%	-0.01	0.03	22%	0.05	0.04
6. Design experiments	30%	0.01	0.02	6%	0.00	0.04	3%	0.08	0.05
7. Class debate	33%	0.02	0.02	11%	-0.02	0.03	4%	0.09	0.06
8. Explain relevance of sci for lives	35%	0.00	0.03	29%	0.01	0.03	19%	0.02	0.03
9. Investigations to test ideas	51%	0.04	0.03	**22%**	**0.07***	**0.03**	8%	0.07	0.05

Notes: All figures can be interpreted in terms of effect sizes, with the 'never' category as the reference group. Percentages refer to the percentage of students within the 'some', 'most' and 'every' group. SE = Standard error. * indicates statistical significance at the five percent level. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included). Full version of questions are as follows: 1= Students are given opportunities to explain their ideas. 2 = Students spend time in the laboratory doing practical experiments. 3 = Students are required to argue about science questions. 4 = Students are asked to draw conclusions from an experiment they have conducted. 5 = The teacher explains how a science idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties). 6 = Students are allowed to design their own experiments. 7 = There is a class debate about investigations. 8 = The teacher clearly explains the relevance of science concepts to our lives. 9 = Students are asked to do an investigation to test ideas.

Table D3b. Estimated association between different types of inquiry-based teaching practices and students GCSE science grades

Guidance Measures	Low Guidance			High guidance		
	Q2	Q3	Q4	Q2	Q3	Q4
The teacher gives students extra help when they need it	0.12	-0.05	-0.01	0.04	**0.08***	0.05
	(0.06)	(0.09)	(0.11)	(0.04)	(0.04)	(0.04)
A whole class discussion takes place with the teacher	0.04	0.05	-0.02	0.12	**0.20***	**0.19***
	(0.04)	(0.04)	(0.06)	(0.07)	(0.06)	(0.06)
The teacher tells me how to improve my performance	0.04	0.06	0.02	0.07	**0.12***	**0.10***
	(0.04)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)
The teacher advises me how to reach my learning goals	0.04	0.03	0.04	0.07	**0.15***	0.10
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)

Notes: All coefficients can be interpreted in terms of effect sizes, with the lowest discovery quartile as the reference group. Standard errors are shown in parentheses. Bold coefficients with a * indicate $p<0.05$. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included). Q2, Q3 ad Q4 refers to quartiles of the inquiry teaching scale.

Table D4. The estimated impact of inquiry-based teaching practices for different sub-groups

		Second quartile		Third quartile		Top quartile	
	N	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Gender							
Girls	2,125	-0.02	0.03	0.00	0.03	-0.01	0.05
Boys	2,236	-0.01	0.04	0.00	0.04	0.05	0.04
Socio-economic status							
Low SES	1,279	0.07	0.05	0.06	0.05	0.06	0.05
Average SES	1,404	-0.03	0.05	-0.06	0.05	0.00	0.04
High SES	1,491	-0.02	0.05	0.02	0.05	0.09	0.05
Science achievement							
Low-achieving	1,216	0.01	0.05	-0.01	0.05	-0.03	0.05
Average-achieving	1,473	0.00	0.05	0.00	0.05	0.07	0.05
High-achieving	1,672	-0.02	0.05	0.02	0.05	0.01	0.06

Notes: All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. * indicates statistical significance at the five percent level. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included). Science-achievement groups based upon top third, middle third and bottom third of pupils in England on the PISA science scale (using the first plausible value). Socio-economic status (SES) based upon thirds of the ESCS index (where data available).

Table D5. The estimated impact of inquiry-based teaching practices for schools with different disciplinary climates

		Second quartile		Third quartile		Top quartile	
	N	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Science class discipline (pupil report)							
Poor discipline	1,373	0.02	0.05	0.05	0.06	0.04	0.06
Average discipline	1,366	-0.09	0.05	-0.05	0.05	-0.02	0.05
Good-discipline	1,358	0.00	0.05	0.02	0.04	0.05	0.05
Science class discipline (School-average report)							
Poor discipline	1,332	0.05	0.05	0.06	0.05	0.03	0.06
Average discipline	1,396	-0.03	0.03	-0.07	0.04	0.05	0.05
Good-discipline	1,369	-0.03	0.04	0.04	0.04	0.04	0.05

Notes: All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. * indicates statistical significance at the five percent level. Students/schools have been divided into three groups, based upon students' reports of the disciplinary climate within their science classes. This is based upon the PISA science 'discipline' scale. Top panel refers to results where students have been divided into thirds based upon their own reports. Bottom panel is where we have used the school average of the discipline scale to divide pupils into groups. Estimates all based upon model specification 4 (see notes to Table 2 for further details on controls included).

Appendix D6. Alternative estimates based upon school-average values of the disciplinary climate scale

	Model 1		Model 2		Model 3		Model 4	
	Effect	**SE**	**Effect**	**SE**	**Effect**	**SE**	**Effect**	**SE**
Inquiry-teaching scale	0.00	0.02	**0.04***	0.01	**0.04***	0.01	0.02	0.01
Observations	**4,361**		**4,361**		**4,361**		**4,361**	
Controls								
Demographics	Yes		Yes		Yes		Yes	
Key Stage 2 scores	-		Yes		Yes		Yes	
PISA science scores	-		Yes		Yes		Yes	
Science subjects studied	-		Yes		Yes		Yes	
Science study minutes	-		-		-		Yes	
Sense of belonging	-		-		-		Yes	
Test anxiety	-		-		-		Yes	
Parent emotional support	-		-		-		Yes	
Before school activities	-		-		-		Yes	
After school activities	-		-		-		Yes	
Perception teacher fairness	-		-		-		Yes	

Notes: The inquiry-based teaching scale is now based upon the average within the school. It has been entered into the model as a continuous term. Hence estimates refer to standard deviation increases in GCSE science scores per standard deviation increase in the (school-level) inquiry-based teaching scale. All figures in the 'effect' column can be interpreted in terms of effect sizes. N = Number of observations and SE = the standard error. * indicates statistical significance at the five percent level.

Online supplementary material E. The estimated association between different types of inquiry-based teaching practices and students GCSE science grades

	Some			Most			Every		
	%	Effect	SE	%	Effect	SE	%	Effect	SE
Items									
1. Explain ideas	20%	-0.03	0.05	38%	-0.03	0.04	37%	0.00	0.05
2. Practical experiments	62%	0.04	0.03	**15%**	**0.09***	**0.04**	4%	0.06	0.06
3. Argue about science questions	37%	0.02	0.02	12%	-0.04	0.03	5%	0.05	0.05
4. Conclusions from experiments	45%	0.05	0.04	36%	0.07	0.04	12%	0.05	0.04
6. Design experiments	30%	0.01	0.02	6%	0.00	0.04	3%	0.08	0.05
7. Class debate	33%	0.02	0.02	11%	-0.02	0.03	4%	0.09	0.06
9. Investigations to test ideas	51%	0.04	0.03	**22%**	**0.07***	**0.03**	8%	0.07	0.05

Notes: All figures can be interpreted in terms of effect sizes, with the 'never' category as the reference group. Percentages refer to the percentage of students within the 'some', 'most' and 'every' group. SE = Standard error. * indicates statistical significance at the five percent level. Estimates all based upon model specification 4 (see Table 2 in main body of the paper for further details on controls included). Full version of questions are as follows: 1= Students are given opportunities to explain their ideas. 2 = Students spend time in the laboratory doing practical experiments. 3 = Students are required to argue about science questions. 4 = Students are asked to draw conclusions from an experiment they have conducted. 6 = Students are allowed to design their own experiments. 7 = There is a class debate about investigations. 9 = Students are asked to do an investigation to test ideas. Items 5 and 8 from the original PISA IBTEACH scale has been excluded; see Table 1 and the data section.

Online supplementary material F. The link between teacher guidance, inquiry instruction and pupil's GCSE science grades

ST100

	Never/Some lessons						Most/Every lesson					
IBTEACH quartile	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**
The teacher shows an interest in every student's learning.	**0.13***	0.06	0.02	0.08	0.07	0.09	0.00	0.04	0.02	0.04	-0.01	0.04
The teacher gives extra help when students need it.	**0.17***	0.07	0.09	0.10	-0.10	0.12	0.02	0.04	0.04	0.04	0.02	0.04
The teacher helps students with their learning.	0.15	0.09	-0.09	0.11	-0.12	0.14	0.01	0.03	0.03	0.03	0.00	0.04
The teacher continues teaching until the students understand.	0.03	0.06	0.03	0.08	-0.11	0.08	0.02	0.04	0.02	0.04	-0.01	0.04
The teacher gives students an opportunity to express opinions.	0.05	0.05	0.04	0.05	0.00	0.08	0.06	0.05	0.08	0.05	0.07	0.05

ST103

	Never/Some lessons						Most/Every lesson					
IBTEACH quartile	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**
The teacher explains scientific ideas.	0.03	0.05	0.02	0.05	-0.04	0.07	0.02	0.04	0.01	0.05	-0.01	0.04
A whole class discussion takes place with the teacher.	0.02	0.04	0.01	0.04	-0.02	0.05	0.09	0.07	**0.16***	**0.07**	0.12	0.06
The teacher discusses our questions.	0.07	0.05	-0.03	0.04	-0.04	0.06	0.02	0.04	0.05	0.05	0.01	0.05
The teacher demonstrates an idea.	0.00	0.05	0.00	0.05	-0.06	0.07	0.07	0.05	0.08	0.05	0.05	0.05

ST104

	Never/Some lessons						Most/Every lesson					
IBTEACH quartile	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**	**Q2**	**Q2 SE**	**Q3**	**Q3 SE**	**Q4**	**Q4 SE**
The teacher tells me how I am performing in this course	0.01	0.04	-0.02	0.04	-0.02	0.05	0.10	0.06	**0.16***	**0.06**	**0.12***	**0.05**
The teacher gives me feedback on my strengths in this subject	0.00	0.04	0.00	0.04	-0.03	0.05	0.08	0.05	0.07	0.06	0.04	0.06
The teacher tells me in which areas I can still improve	0.03	0.04	0.02	0.04	0.01	0.06	0.07	0.04	**0.11***	**0.05**	0.08	0.05
The teacher tells me how I can improve my performance.	0.02	0.04	0.03	0.04	0.03	0.05	0.07	0.05	0.09	0.05	0.06	0.05
The teacher advises me on how to reach my learning goals	0.03	0.04	0.01	0.04	0.02	0.05	0.09	0.05	**0.12***	**0.05**	0.08	0.05