A developmental evaluation approach to lesson study: exploring the impact of lesson study in London schools.

Corresponding author

Dr David Godfrey
Department of Learning and Leadership
UCL Institute of Education, London
Bedford Way
London WC1H 0AL
David.godfrey@ucl.ac.uk

Co –authors

Sarah Seleznyov
Bespoke Programme Leader
UCL Institute of Education
s.seleznyov@ucl.ac.uk

Dr Jake Anders
Senior Research Associate
UCL Institute of Education
jake.anders@ucl.ac.uk

Nicholas Wollaston
Research Officer
UCL Institute of Education
nicholas.wollaston@ucl.ac.uk

Dr Fabián Barrera-Pedemonte
Lecturer in Educational Psychology, Universidad del Desarrollo, Chile
Research Fellow, CIAE, CIDER and OECD
fabian.pedemonte.14@ucl.ac.uk

Acknowledgements

Many thanks to the hard work of the teachers and school leaders involved in the 'Ascend' and 'Lambeth Connecting Knowledge' LSEF projects about which this research describes.

A developmental evaluation approach to lesson study: exploring the impact of lesson study in London schools.

Abstract

This paper presents a methodology for the developmental evaluation of a lesson study programme in primary and secondary schools. Our approach combined the principles of, i) user-focused evaluation, in which as evaluators we acted as participatory members of the innovation team and sought to involve users in the design and implementation of evaluation tools, ii) a multi-level logical model to guide data collection and impact measurement, and iii) an 'improving rather than proving' approach to evaluation. The evaluation tools were used on a programme to promote lesson study in London schools involving 133 teachers and 33 schools. The evaluation methodology included outcomes at school leadership, teacher and student levels. Issues of internal and external validity are discussed and strengths and weaknesses are described. Findings showed promise in the use of our scale to measure changes in teacher pedagogical outcomes and in the recording of qualitative changes to both teachers and students as a result of the lesson study cycles. Suggestions for the future use and development of this methodology are proposed, including better use of control groups and quantitative measures to record changes in learning outcomes for students.

Keywords: Lesson study; evaluation; Guskey; professional development

Introduction

Lesson study is an approach to teacher professional development from Japan involving collaborative planning of a lesson, live lesson observation and reflective discussion. Lesson study has been increasingly adopted worldwide as a form of teacher learning and school improvement. This paper raises questions about the evidence base for lesson study thus far and proposes a new developmental evaluation (Patton, 2002) approach to exploring the impact of lesson study on teachers and pupils, contributing to the literature on the evaluation of teacher professional development (PD).

Through this model we seek to show how evaluation, when 'built-in' from the start, has the potential to build focus and coherence to teachers' learning, maximising the kind of high quality professional discussion and skills development that leads to positive outcomes for pupils. Throughout, we have balanced using a practical approach that schools can adopt relatively easily, with being sufficiently rigorous to be able to identity the features of the programme that led to its successes and failings, and to satisfy the requirements of those needing to know that the investment of time and resources has had an impact.

We present a systematic approach to the evaluation of the impact of the lesson study, used in a programme involving 133 teachers and 33 London schools. We also invite further debate and research on ways in which lesson study can be evaluated.

This article reviews the current empirical evidence on lesson study and looks at how many studies show a logical chain of impact from teacher learning to student outcomes. We then examine a lesson study programme the followed a developmental evaluation approach that sought to address this sequence of impact. Throughout the analysis, and in our conclusion, while outlining some of the impact of the programme, we focus on the strengths and limitations of the evaluation tools we used and offer some suggestions for future research following a similar developmental evaluation methodology.

Before that, the sections below briefly outline the context and details of the programme itself.

Lesson study in England

Lesson study, a professional development model from Japan, has become increasingly popular in the last decade or so as a form of professional development in English schools (Hammersley-Fletcher et al, 2015, Lewis, 2006, Barber, 2007). Lesson study has emerged in the context of a global race in which countries (or cities) seek to increase pupil attainment (e.g. Sahlberg, 2011). This race has led to an increase in what Steiner-Khamsi (2012) describes as 'travelling reforms', reforms borrowed from other nations that are perceived to be models of excellent practice. The Far East, and Japan, Singapore and China in particular have seen several key aspects of their educational systems lauded internationally (Moore, 2014).

One of the dangers of borrowing ideas from elsewhere is the 'dilution' of those aspects of practices that make them particularly successful in originating countries. As HE facilitators, programme co-designers and evaluators, one of our concerns was to ensure fidelity to the key features of lesson study in the London schools' context, while allowing for flexibility and fitness for purpose**.** The distinctive features of lesson study have been written about in detail elsewhere (see Takahashi and McDougal, 2016 and Stigler and Hiebert, 2009 for example). We summarise the essential elements of lesson study from this literature as involving:

1. identification of a research focus through an analysis of 'data' (in its most inclusive definition);
2. collaborative and detailed planning of a 'research lesson' addressing the agreed research focus so that the lesson has shared ownership;
3. a live research lesson, in which one member of the planning group teaches the lesson and all teachers in the group observe pupil reactions;
4. a post-lesson discussion to draw out learning in relation to the research focus.

The approach to lesson study we describe below, adopted all the above features.

The lesson study programme on which this research focuses

Between 2013 and 2015 the London Schools' Education Fund funded over 100 teacher professional development projects in the Greater London metropolitan area,[1] including several lesson study projects. The project from which the empirical data was taken was concerned with improving outcomes for learners, primarily via engaging teachers in lesson study. Involving both primary and secondary school teachers, this was targeted at under-achieving or vulnerable pupil groups in either Mathematics or English. Pupils were ethnically diverse with a wide range of minority groups, as might be expected of their urban location. A high proportion of pupils were eligible for free school meals and many were in receipt of support for special educational needs.

The data in this study is taken from the second year of the intervention, as the first year's project evaluation, which was conducted and explored with project stakeholders, led to changes in the programme in year two and improved evaluation tools to better meet the needs of the schools and teachers. The pupil cohorts are also separate from year one to year two of the project, thus here we report only on the pupil outcomes from the 2014-15 cohort: 133 teachers and 33 schools: 26 primary and 7 secondary. 334 pupils were selected as target pupils; with most teachers selecting 3 from one of their classes.

Teachers on the project participated in several lesson study cycles of enquiry. In each cycle, one teacher would be the focus of a cycle and work alongside colleagues to co-plan a lesson, then the focus teacher would teach the lesson to their class, and colleagues would observe the learning of case study pupils (between 3-6), followed by a de-briefing session. In the post lesson discussion, the lessons learned would be used to plan a second cycle of planning, teaching and discussion with a different member of the group acting as focus teacher.

[1] The London Schools' Excellence Fund was funded jointly by the Greater London Authority and England's Department for Education across two years and schools or other educational organisations were able to apply for up to £250,000 to support teacher professional development: https://www.london.gov.uk/what-we-do/education-and-youth/improving-standards-schools-and-teaching/london-schools-excellence?source=vanityurl

Previous research on the impact of lesson study

In a systematic search of the international literature, we sought to review the breadth of evidence for the impact of lesson study through recent empirical studies.
Searches were conducted in the British Education Index, Education Resources Information Centre and Australian Education Index. We found 210 empirical studies on the impact of lesson study published between 2005 and 2017. 24 were additionally located through checking other recent systematic reviews of lesson study (Saito 2012; Yee Wong and Ming Cheung, 2014; Xu and Pedder 2014). Of these, 174 were rejected because they:

- were not about in-service teachers
- focused on only one aspect of lesson study, for example discourse analysis of teacher talk
- did not include live observations of lessons or the identification of a research focus
- described a hybrid form of lesson study, for example Learning Study or Design Study as identified by Xu and Pedder (2014)
- included no empirical data on the impact of lesson study.

Appendix two has the full list of the 36 included studies; 20 took place in the USA and 7 in the UK, with others taking place in Australia, Canada, The Netherlands, Indonesia, Japan, Philippines and Singapore. In terms of sample sizes, numbers of teachers ranged from 3 to 83. The studies are largely small-scale ones involving small numbers of teachers and schools; the average number of teachers involved was 15, 24 involved 5 schools or fewer and 16 studies involved only 1 school.

In our review we were interested in the extent to which evidence was presented in prior empirical research about the entire logical chain of impact of lesson study (as a form of professional development). For this we relied particularly on Guskey's (2000) model which suggests five levels at which the effectiveness of PD can be measured. Following advice by Lewis (2000) and Fernandez and Yoshida (2004), we identified several anticipated outcomes for lesson study in the English context:

1. Teachers' reactions

Teachers' attitudes to and enjoyment of professional learning might improve.

2 Teachers' professional learning

There could be an impact on teacher confidence and the quality of teacher professional dialogue through the collaborative elements of LS.

3. The organisation's professional development model

The structure, time, resourcing of the school's professional learning programme would need flex in order to accommodate lesson study. Cultural attitudes towards professional learning might shift for example in relation to teacher ownership of learning, and lesson observation as learning not performance.

4. Teacher use of new knowledge and skills

Teachers' newly acquired confidence, pedagogical content knowledge and knowledge of their pupils could lead to changes in practice.

5. Pupil learning outcomes

Changes in teachers' practice might lead to improved learning for pupils.

Our analysis of the 36 empirical studies found:

- 21 studies analysed teachers' reactions to and enjoyment of the lesson study process (outcome 1).
- All 36 studies analysed changes caused by lesson study in relation to teachers' learning (outcome 2), including one or several of the following aspects: subject content knowledge, pedagogical content knowledge, teacher confidence.
- Only 6 studies considered the impact the lesson study process had had on the schools' professional development model (outcome 3).
- Similarly, only 17 studies considered whether teachers had changed their practice (outcome 4) in response to their involvement in lesson study. The literature describes lesson study as a "comprehensive system for teacher learning" (Perry and Lewis, 2009, pp. 19), and yet 64% of studies did not explore any evidence of transfer of learning from lesson study to teachers' general classroom practice.
- Only 12 studies considered the impact the lesson study process had had on pupil learning (attitudes and/or progress) (outcome 5), the majority of these studies relying on teacher reports of improvements to learning, rather than standardised measurement tools such as questionnaires or tests.
- Overall, only three studies captured evidence about all the above five levels.

In summary, we sought to address some of the limitations of previous research on lesson study by applying an evaluative model that would discriminate effects at each level of impact. While our concern was not primarily to demonstrate the overall efficacy of lesson study in comparison with other approaches (or a control group), the discussion of our developmental evaluative approach is nevertheless applied to a very large sample of teachers and schools, compared to previous empirical work.

Our developmental evaluation methodology

Our approach aligns particularly with Patton's 'developmental evaluation' (2010) and Cousin's 'practical participatory evaluation' (Cousins & Earl, 1995), in that we sought to engage teachers - as primary users of lesson study - actively and directly in all stages of the evaluation. Overall, in terms of Christie and Alkin's evaluation theory tree (2013), we see this as 'use-focused evaluation'. Thus, we focused less on concerns about methods for collecting the most rigorous knowledge (the methods branch) or on issues surrounding the subjective valuing process (valuing branch), rather our evaluation was designed to aid decision-making and changes to practice by users (the use branch).

The project was supported by staff working from the London Centre for Leadership in Learning at UCL Institute of Education, primarily by two of the authors of this paper. As evaluators, we acted as participatory members of the design team for the project rather than separating ourselves from the project as evaluators (Patton, 2010). We were involved, as much as possible, as HE advisers in the co-creation of the programmes and in its modifications and developments for year two and beyond. We co-planned project logistics, facilitated lesson study training, and co-designed and implementing process and outcome evaluation.

For teachers involved in lesson study trios, there were three elements to our evaluation approach: i) setting specific final impact goals for case pupils and own practice, ii) the planning and conduct of lesson study cycles and iii) evaluation of LS cycles by the focus teacher and also of overall learning from all participants at the end of the year.

This interactive process aimed to empower teachers as stakeholders in the project (Fetterman, 1996), this process having the parallel intention of developing the skillset of teachers to evaluate their own future lesson study projects. Our partnership role would also let us explore the lesson study process with the deliberate intention of informing and improving its implementation (Cousins & Earl, 1995). Our ongoing interaction with the process of the

project, its design team and its primary users enabled us to operate a series of feedback loops with the intention to maximise the effectiveness of the project.

Our approach to the impact evaluation of lesson study included each element of the lesson study cycle and the experiences of both focus and collaborating teachers. Below we analyse the methodology, tools and results for our evaluation of the impact of lesson study in this project. We also sought to establish the logical chain between teacher learning and pupil outcomes based closely on our interpretation of Guskey's (2000) levels (see above). Our evaluation tools sought to address each of these levels. A summary of these instruments, their aims and the response rates can be found in Table 1. The six sections that follow analyse our use of each of the evaluation tools listed in the left hand column of the table.

Table 1: Summary of data collection tools and response rates

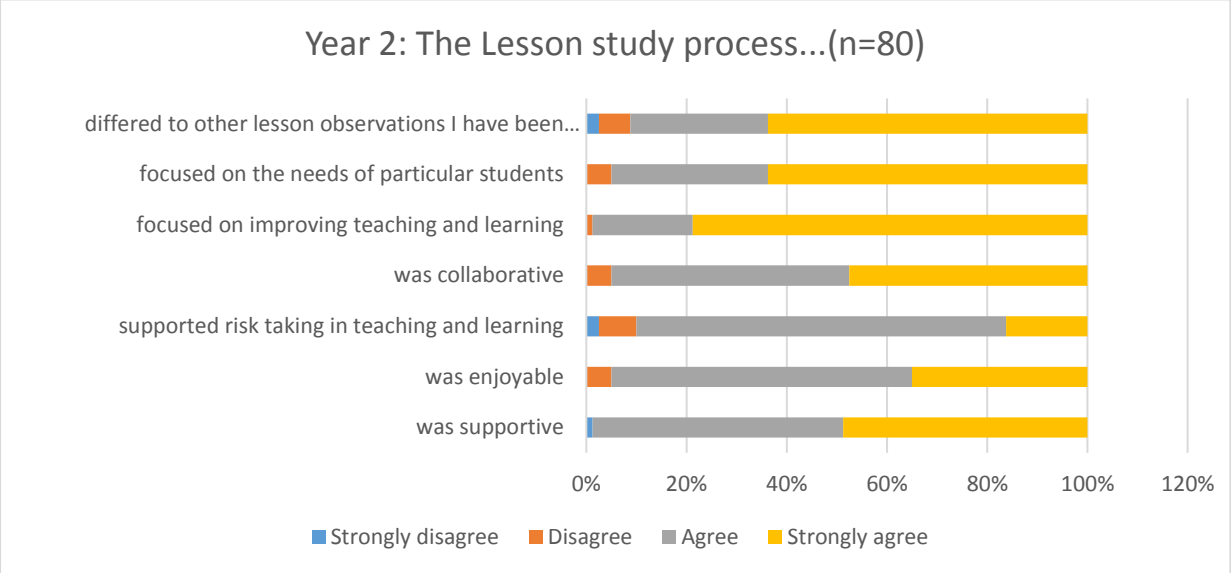
Data collection tool	Level of impact (Guskey)	Return rates (133 teachers, 334 pupils in total)	Metric used
Overall lesson study survey	1. Teachers' reactions: attitudes and enjoyment	78 teachers completed online survey	3 open ended questions on teacher learning and 8 Likert Scale responses (4 point scale Strongly Agree-Strongly Disagree) on reactions to LS process
Teacher self-evaluation surveys	2. Teacher professional learning	72 teachers completed baseline and final self-evaluations	Nine item scale of 1-7 for each (7 highest). Baseline and end of year evaluations
Interviews: school leaders and teachers in a sample of schools	3. Organisational support and change	3 case schools visited	Semi-structured interview prompts
Interviews: focus group of teachers leading lesson study groups	3. Organisational support and change	6 Cohort 1 teachers interviewed in group	Open responses in group interview on lessons learned about leading LS
Impact framework tool: written analysis and self-evaluation using	4. Changes to teacher use of knowledge and skills 5. Changes to pupil learning outcomes (qualitative)	53 teachers carried out impact assessments (relating to over 160 target pupils)	Teachers record baseline and impact descriptions of teaching practice
Focus teacher surveys	5. Pupil learning outcomes (qualitative)	Data from 66 teachers and 221 case study pupils	Record of observations about pupils and assessment of extent and type of improvements to learning on 5 point scale
Quantitative changes to pupil progress	5. Pupil learning outcomes (quantitative)	334 pupils	Achievement of target level (set by teachers at start of year)

Analysis of the use of developmental evaluation in lesson study

i) Overall lesson study survey

In order to capture the key professional learning features of LS cycles over the course of the year and the dual experiences of focus and collaborator teachers, we captured responses to seven statements on a four point Likert scale and one open ended invitation for comments. The results can be seen below.

Figure 1: Reflections on professional learning experience throughout the year



Most participants felt that the process was collaborative, focused on the needs of their pupils and on improving teaching and learning. Weaker agreement was towards risk taking; comments reflected the importance of good relationships and participants felt that risk taking grew once trust was built up in the group.

The focus on pupils was seen as key; professional learning was largely seen as a product of the efforts made to improve pupil learning, this was the key motivation too. By focusing observations on pupils, the teacher also felt less threatened and judged.

ii) Teacher self-evaluation survey

Teachers were asked to self-assess their confidence on nine areas of pedagogy:

1. Consideration to pupil voice
2. Understanding of the pedagogic process
3. Clear thinking about longer term learning outcomes
4. Building on pupils' prior learning and experience
5. Scaffolding pupil learning
6. Using a range of techniques
7. Developing higher order thinking and metacognition
8. Embedding assessment for learning
9. Inclusivity

These nine areas were based on a report by Husbands and Pearce (2012) that reviewed a range of literature for the then National College for School Leadership[2]. This was converted into a baseline and end of year self-assessment. All teachers taking part in the LS cycles were urged to complete these measures as a way of charting their professional learning and changes to perceived efficacy in areas of pedagogy. This served two purposes. First, we aimed to provide a more rounded view of pedagogical standards than teachers (influenced by the dominant influence of Ofsted[3] 'good' and 'outstanding' criteria) may be accustomed to doing. The second purpose was to encourage self-evaluation by teachers, who could then think about the skills and knowledge they wanted to develop over the year, particularly in relation to the aims they had set out for their case pupils.

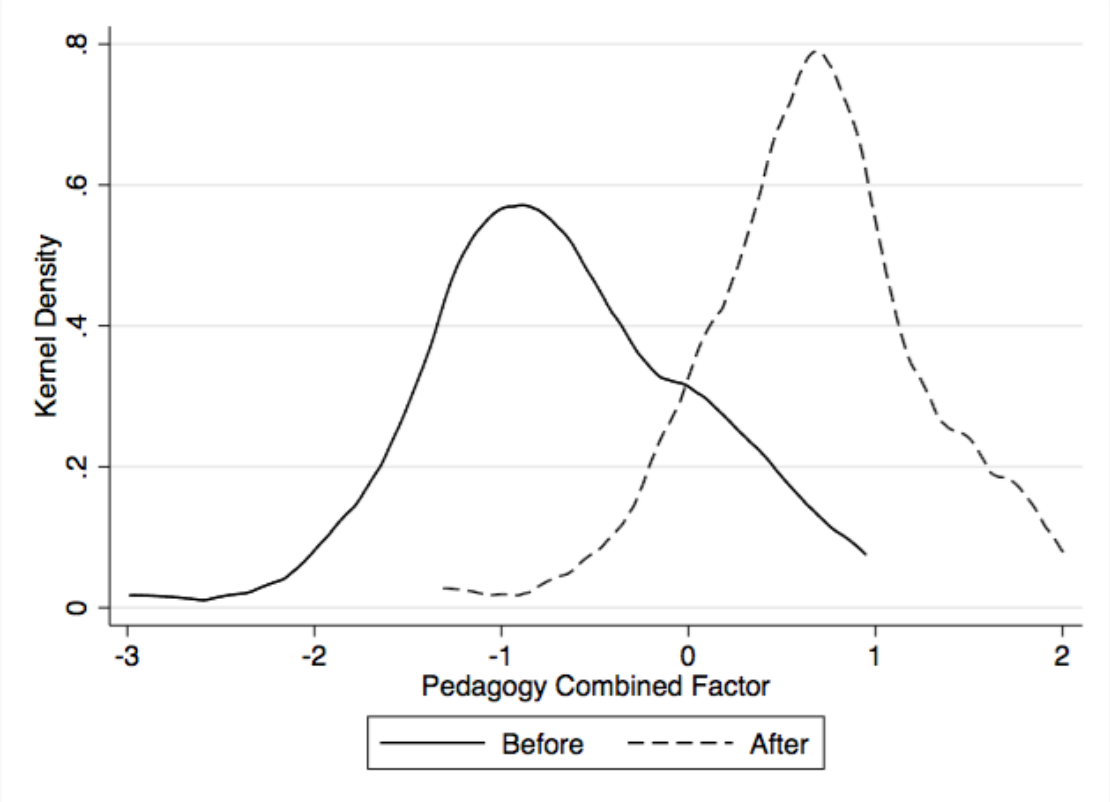
Tests of internal validity were conducted on the use of the teacher self-evaluation tool using factor analysis. This was carried out using the sample pooled across both time points and

[2] Now re-named the National College for Teaching and Leadership.
https://www.gov.uk/government/organisations/national-college-for-teaching-and-leadership
[3] Office for Standards in Education is the external inspectorate for state schools in England

combined the responses on the nine variables. Based on the moderate positive correlations observed between these items, a single factor which explained 53% of the total variance was extracted. The composite measure had meaningful factor loadings for all nine items (all greater than 0.5). These results suggested that the self-evaluation tool was assessing a single construct related to high-quality pedagogy; in other words, the higher the scores in all these nine items, the higher the values in the underlying factor score. It also suggests that we can use our extracted factor as a summary measure of high-quality pedagogy to consider how this has changed during the period of the intervention.

Despite being based on a pooled factor analysis, plotting the distribution of this factor reveals that it is bimodal. Plotting the distribution separately by time point (Figure 3) reveals two unimodal distributions with the post-test distribution significantly more positive and noticeably less spread than the pre-test distribution. We can also summarise the change in this factor using means, which we report in Table 2.

Figure 2: Distribution of pedagogy factor score before and after the intervention



Notes: Kernel density estimate of the distribution of pedagogy combined factor plotted separately by time point.

Table 2: Teacher self-evaluated self-efficacy scores pre- and post-test using a composite pedagogy factor based on the nine

Corresponding data (only when there is exact pairing)	**Composite pedagogy factor**
Pre-test	-0.67
Post-test	0.67
Raw difference	**1.34**
Standardised difference	**2.0**

Notes: Sample: Teachers responding at both time points. Pre-test and post- reports average factor score at respective time points. Standardised difference calculated by dividing mean difference between pre- and post-tests by pooled standard deviation (i.e. analogous to Cohen's d effect size).

As shown in Table 3, seemingly large (and, in all cases, statistically significant) changes in each of the nine pedagogical areas were found, although it is unclear against what baseline sizes should be judged. We can compare these changes against one another by converting these to standardised differences (i.e. dividing the raw difference by the pooled standard deviation of the measures), noting larger increases in teachers' self-reported self-efficacy in "consideration to pupil voice" and "building on pupils' prior learning and experience" but smaller increases in "inclusivity", "clear thinking about longer-term learning outcomes" and "developing higher order thinking and meta-cognition".

Table 3: Teacher self-evaluated self-efficacy scores pre- and post-test on nine pedagogical areas

Corresponding data (only when there is exact pairing)	1) Consideration to pupil voice.	2) Understanding of the pedagogical process	3) Clear thinking about longer term learning outcomes	4) Building on pupils' prior learning and experience	5) Scaffolding pupil learning	6) Using a range of techniques,	7) Developing higher order thinking and meta-cognition	8) Embedding assessment for learning	9) Inclusivity
Pre-test	3.5	4.6	4.9	4.8	4.7	4.9	4.6	4.6	5.0
Post-test	5.3	5.7	5.8	6.0	5.8	5.9	5.7	5.7	5.9
Raw Difference	**1.8**	**1.1**	**0.9**	**1.2**	**1.1**	**1.0**	**1.1**	**1.1**	**0.9**
Standardised difference	**1.7**	**1.3**	**1.0**	**1.4**	**1.2**	**1.2**	**1.0**	**1.2**	**1.0**

Notes: Sample: Teachers responding at both time points. Pre-test and post- reports average raw score from each measure at respective time points. Standardised difference calculated by dividing mean difference between pre- and post-tests by pooled standard deviation (i.e. analogous to Cohen's d effect size).

iii) Interviews

We interviewed headteachers and other senior leaders, as well as participating teachers in situ, at three participating schools. We wanted to find out about the pedagogical leadership skills required to maximise the impact of LS at middle and senior leadership levels.

Interviews revealed concerns about the consistency with which school leaders were supporting the efforts of teachers to work collaboratively in LS cycles over the school year. Key to the success of LS was perceived to be protected time for teachers to work together and leadership management of LS cycles. In some schools teachers were asked to find time after school or other 'free' periods to discuss the lesson they had planned together and observed. Where LS cycles were more effective, school leaders helped coordinate who would work with whom, and when to conduct each phase of the LS process. Protecting teachers from other responsibilities was key to this. Some participating teachers were also removed from the requirement to be observed as part of a performance management review. This enabled trios to work together in full trust and to take risks, thus focusing on learning from peers, rather than 'proving' competency to their seniors. These interviews were useful at the interim stage of the project to fine tune our advice to teachers and also to provide guidance to senior leaders in how to get the most from lesson study. These were also important to determine the factors that would ensure successful implementation beyond the two year funded period.

iv) Impact framework tool

At the start of the academic year (around October) teachers were asked to identify between 3 and 6 case pupils in their classes whose learning required support and to identify a specific aspect of their learning needs, for instance problem solving in mathematics. This did not mean that the lesson study process would not benefit all pupils in the class, but enabled teachers to prioritise pupils whose learning needs were greater. For these pupils, they were asked to imagine what successful learning would look like by the end of the academic year (July), and to make qualitative statements that described what pupils would be saying, doing and feeling in relation to this focus. Teachers had to avoid using general terms such as 'the pupil will be motivated', instead, they might for example write, 'pupil x remains on task for at least 20 minutes of independent working time without adult prompting'. We also discouraged the use of jargon terms like 'peer learning', instead using simple and specific language, such as, 'pupil x usually asks his partner for help to select an appropriate strategy to solve mathematical problems'. Similarly, we asked for statements about their desired teaching practice in relation to these focus pupils. The next step was to ask teachers to describe the *baseline situation*: describing pupils' current learning and their own teaching practices in relation to the focus. These baseline-impact descriptions were called Impact Frameworks. At the end of the year, we asked teachers to colour-code each of their impact statements, to show red: not achieved, amber: partially achieved or green: fully achieved ('RAG' rating). These gave us a rich source of information about the types of impact that teachers achieved with their pupils over the course of the year.

Qualitative analysis of 155 Impact Frameworks from 53 teachers was carried out using NVivo (N = 53) on the colour coding given by teachers to each of their statements.

Changes in practice described by participants were analysed using the same nine categories as used in the teacher self-evaluations, together with an additional category for 'other teacher competences'. 108 (70%) of the statements were categorised as either Embedding Assessment for Learning (23%), Inclusivity (24%) or Scaffolding Pupil Learning (23%). No impact statements were coded as 'Clear Thinking about Longer Term Outcomes', nor as 'Understanding the Pedagogical Process'; this is an interesting finding since these areas were still perceived as significant areas of development in the teacher self-evaluations. Presumably

these were seen as indirect areas of learning as a result of the more specific changes to practice identified above.

There were 468 impact statements describing the impact of the lesson study project on the pupils' experience and learning. 86 (18%) of these statements were categorised as an area of the national curriculum in England known as 'Number', 95 statements (20%) were coded as either Problem-solving in Mathematics or the Use of Spoken Language in Mathematics. Of the 86 impact statements in Number, around half were achieved, over a third were partially achieved, whilst only 9 were not achieved.

Of those referring to literacy aims, 58 (12%) of the statements were categorised as Writing, whilst 29 (6%) were shared between Reading and Spoken Language. Of the 58 impact statements in Writing, almost two-thirds were achieved, almost a third were partially achieved, whilst only 5 were not achieved.

There were also a number of non-subject specific aspects of learning that are interesting to report on: 127 impact statements (27%) mentioned Collaboration, Confidence, Perseverance and Resilience and Engagement (16%). Of the 73 impact statements in Engagement, over half were achieved, over a third were partially achieved, whilst only 3 were not achieved.

Teachers reported the value of the impact frameworks in both helping to plan the lesson for the LS cycle and providing specific observational foci for colleagues. The colour coding was also motivational to individual teachers as it enabled them to get a visible sense of their own success. Moreover, evaluating the project as a whole, the RAG ratings gave a sense of the (albeit self-reported) success of teachers' approaches over the year with their target pupils. Of changes to practice statements, 65% were felt to be achieved, 30% partially achieved and 5% not achieved. On changes to pupil outcomes, these were 57%, 36% and 7% respectively.

v) Focus teacher survey

The focus teacher for each lesson study cycle would refer to their Impact Framework when planning a lesson (see section IV, above). The Impact Framework provided a useful overall focus, within which the lesson plan would have a subordinate linked aim. The lesson planning was often related to a shared focus between the three teachers. Planning also involved discussion about the context of the class and in particular about the case pupils.

After the second lesson and post lesson discussion, the focus teacher would take detailed notes and then use these to complete the focus teacher questionnaire. For each example of pupil learning noted, observers also rated the extent of this change on a four point scale: no change; a change to a specific aspect of learning in the subject; improvement which should impact on pupils' achievement in this subject; or profound transformation to this pupil's learning in this subject. Table 4 below, shows the distribution of these individual entries on this scale. Note, more than one 'entry'/observation' was possible for one case study student.

Table 4: The extent of improvements to case pupil learning during a lesson study cycle:

No. of entries	Extent of change in learning by case study pupils		
Total	No. of entries	%	Level of change
301	40	13%	No change (or worse than expected)
	129	43%	Improvement to a specific aspect of learning in this subject
	120	40%	Improvement which should impact on pupils' achievement in this subject
	12	4%	Profound transformation to this pupil's learning in this subject

Perhaps unsurprisingly, the majority of statements related to improvements to specific aspects of learning and those that could have a lasting impact on subject knowledge, with only a small percentage relating to profound transformations. The fact that 13% of observations showed no improvement or worsening again should not be too surprising during one single lesson study cycle. When these observations were coded into specific themes we found that the top observations concerned: Increase in understanding/usage of knowledge (17%); Increase in engagement/participation (16%); Positive impact on confidence (14%). An interesting area for development here may be in how teachers can look for key aspects of the pupils' learning that are likely to lead to the biggest changes in their progress over time and to see the LS cycle as an exercise in gaining insight to these aspects.

Some teachers made entries that may reveal a misunderstanding about what was required here; an example of this being ‘increased attendance’. Another entry, ‘changing of seating’ was more about the cause of the improvement in learning rather than the outcome. Such comments highlighted the need for clear exemplification and guidance on the use of this evaluation instrument.

vi) Quantitative changes to pupil progress

This aspect was the most challenging for several reasons. Prior to starting year two of the projects, the government had decided that schools should no longer used standardised national measures of progress. This has become known as ‘assessment without levels’ (McIntosh, 2015). This meant that we no longer had a measure of progress that all school participants could agree on across the projects. Our approach to this was to say that we would use a scale determined by each teacher, which would allow for them to say if a target was achieved for each of their case pupils. Essentially this meant that the teacher would note the baseline achievement level for their case pupils, measure progress for a few months, and then project this rate of progress to the end of the year. They would then try to outstrip this projection, setting an aspirational target. Unfortunately the implementation of these projected targets was inconsistent. While there were significant positive statistical outcomes, we do not consider these ‘safe’; in some cases there was a misunderstanding in the application by teachers. We were also highly sceptical that one year of data would be able to reliably show increases in pupil attainment using normal school test and progress measures, especially if these were not set against a matched sample of pupil data. We nevertheless retain the notion of a quantitative measure of changes to pupil’s learning in our model as it could be a clear indicator of the success or otherwise of the approaches. Most appropriate though, would be to design specific tests which measure the progress on aspects of the curriculum that teachers intended to improve during the year. These could then be used to make comparisons across cohorts, schools, year groups etc. Unfortunately these were not used or anticipated in this project.

Discussion

In summary, we feel the model we have presented here is promising as a developmental evaluation process and feel that this could inform the evaluation of other LS projects. We have captured outcomes at all stages of Guskey's model over the course of a year, albeit to varying degrees of success. The lesson study survey helped us understand the features of LS professional learning that were effective, and the degree to which teachers enjoyed and engaged with the LS process. Our teacher self-evaluation measured a single construct for pedagogy and teachers were able to provide distinctive end of year recordings to show the distance they had travelled in their learning. The interim interviews with school leaders provided insight into the challenges of organising LS effectively so that its effectiveness could be maximised. In particular, we showed how school leaders needed to strongly support the time needed to take part in LS cycles. The qualitative outcomes, particularly from the Impact Frameworks, were successful in cohering teacher efforts when working in LS trios; they focused minds on the purpose of each LS cycle and allowed teachers to make sharp observations about the extent of their success at the end of the year. Coding these qualitative outcomes enabled us to look at the overall success rates of teachers in relation to specific pupils; but it was also highly motivating for teachers to see their successes and also to highlight areas to where they still needed to work. Having these feedback loops in the evaluation process, including recording the micro-impact of LS cycles, enabled important adjustments to be made to teaching strategies.

There were also challenges. The quantitative pupil outcomes were unreliable and in any case, may have already lacked validity, since general measures of progress may have been insufficiently tailored to the aims of the projects to pick up short to medium term gains in achievement. Therefore, we could not show whether LS made a distinctive contribution to schools' progress or attainment measures, nor show whether a difference was made to the 'non-case study pupils' in each class.

One solution to this would be to identify an 'object of learning' that would be the focus of teacher's shared aims. For instance, we saw that most of the mathematics projects were on problem-solving aspects of the curriculum. For this, it would make sense to devise specific tests to measure progress in this area. Subject experts could advise on the creation of appropriate tests and thus enable easier judgements to be made about the success of the LS project across different school contexts. Looking at the distinct progress of control classes or

year groups whose teachers were not involved in LS trios, and of the non-case pupils within LS cycles could also be analysed with such tests.

The lack of control groups for the teacher self-evaluation means that it is difficult to know if LS made a distinct contribution to participants' learning other than what they may have expected compared to 'business as usual'. Attrition rates were high for some of the evaluation tools, in particular where communication was diluted across large numbers of schools. We also relied extensively on self-report measures. The emphasis on evaluation to 'improve' rather than 'prove' led to a trade-off. On the one hand, we could have sought standardised measures of improvement, such as using Ofsted (inspection agency) observation grades for teachers – to prove (or not) that lesson study makes teachers 'teach better'. Instead, we chose to define pedagogy in a way that made it much trickier to independently measure but much more useful for teachers to reflect on. Choosing the latter was a conscious decision to measure 'things that matter' (Eisner, 1976). However, while more difficult, this is not logically impossible to measure empirically.

In terms of comparing to 'business as usual', we are conducting trials with other cohorts of teachers, such as trainees in their first year of teaching practice. Using the same scale, we are assessing their pre and post-test scores and this will make it possible to see if LS teachers report more or less gains compared to this sample. We could also capture these data for teachers at different parts of their careers and in different school phases. It may be interesting to see if LS involvement provides a distinct contribution to pedagogy when looking at the nine dimensions of our teacher evaluation scale. It makes most sense for evaluators to give out the self-evaluations to other teachers who may be not involved in LS but may be in the same or a similar school or context.

In our project, there was also a tension between having 'static' desired outcomes and baseline measures and the reality of a dynamic and evolving picture. While the Impact Framework helped immensely to make sense of why teachers were conducting LS cycles, as teachers learned about their case study pupils, they often revised their initial idea of both the baseline and desired impact. This means our Impact Framework ought to have a separate section for 'non-anticipated' or 'evolving' targets and impact on their pupils and practice. For some teacher groups, the use of a learning diary was a really useful additional feature and the sharing of pictures, artefacts and reflections provided further helpful professional learning through discussion.

Finally, our own involvement in this developmental evaluation process revealed the unique contribution that university staff could make. We facilitated leadership of LS within and across schools, consulted with stakeholders, devised and analysed the evaluation tools, fed back mid-project to recommend improvements to the implementation of LS and acted as a critical friend to teachers involved in LS cycles. We found it important to guide teachers on how to formulate the qualitative goals for their impact frameworks. Otherwise, it was all too easy for teachers to make imprecise statements that would allow impact to be observable. We also encouraged teachers to make aspirational targets, and thus pushed them to find innovative ways to reach them with underachieving pupils. Efforts by school leaders to make lesson study sustainable should therefore be mindful of the need to retain these qualities by either continuing to involve external staff or to build capacity internally for this pedagogical leadership to be addressed.

References

Christie, C. and Alkin, M. (2013). An Evaluation Theory Tree. In Alkin, M. (ed.) *Evaluation roots: a wider perspective of theorists' views and influence.* pp.11-57. London: Sage Publications.

Coe, R., Aloisi, C., Higgins, S. and Major, L.E. (2014). What makes great teaching? Review of the underpinning research. Available at: http://www.suttontrust.com/wp-content/uploads/2014/10/What-Makes-Great-Teaching-REPORT.pdf [Accessed on 21.04.17]

Cousins, J. and Earl, L. (eds.) (1995). *Participatory evaluation in education: studies in evaluation use and organizational learning.* London: Falmer Press.

Fernandez, C. and Yoshida, M. (2012). *Lesson study: A Japanese approach to improving mathematics teaching and learning*. New York: Routledge.

Fetterman, D. (1996). Empowerment evaluation: an introduction to theory and practice. In Fetterman, D., Kaftarian, S. and Wandersman. A. (eds.) *Empowerment evaluation: knowledge and tools for self-assessment and accountability.* pp.3-48. Thousand Oaks:, CA: Sage.

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

Hammersley-Fletcher, L., Lewin, C., Davies, C., Duggan, J., Rowley, H., & Spink, E. (2015). Evidence-Based Teaching: Advancing Capability and Capacity for Enquiry in Schools. *Interim report, Nottingham: National College for School Leadership*.

Husbands, C. and Pearce, J. (2012). "What makes great pedagogy? Nine claims from research." *Research and development network major themes: Theme 1*.

Lewis, C. (2006). Lesson study in North America: Progress and challenges. *Lesson study: International perspective on policy and practice*. 7(36).

Lewis, C. (2000) *Lesson Study: The Core of Japanese Professional Development*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

McIntosh, J. (2015). Final report of the Commission on Assessment without Levels, September 2015. (Available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/483058/Commission_on_Assessment_Without_Levels_-_report.pdf (Accessed on, 31/3/17)

Moore, A. (2014) Understanding the School Curriculum: Theory, politics and principles. *London: Routledge.*

Patton, M. (2010). *Developmental evaluation: applying complexity concepts to enhance innovation and use.* New York: Guilford Press.

Perry, R. and Lewis, C. (2009). What is successful adaptation of lesson study in the US? *Journal of Educational Change*, *10*(4), pp.365-391.

Sahlberg, P., 2011. The fourth way of Finland. *Journal of educational change*, *12*(2), pp.173-185.Saito, E. (2012) Key issues of lesson study in Japan and the United States: A literature review. *Professional development in education* 38.5: 777-789.

Steiner-Khamsi, G., and Waldow, F. (eds) (2012). World Yearbook of Education 2012: policy borrowing and lending in education. *London: Routledge*

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

Takahashi, A. and McDougal, T. (2016). Collaborative lesson research: maximizing the impact of lesson study. *ZDM*, *48*(4), pp.513-526.

Yee Wong, W., and Ming Cheung, W. (2014). Does lesson study work? A systematic review on the effects of lesson study and learning study on teachers and students. *International Journal for Lesson and Learning Studies*, *3*(2), pp.137-149.

Xu, H. and Pedder, D. (2014) Lesson study: and international review of the research. in Dudley, P. (ed.) (2014). *Lesson Study: Professional Learning for Our Time*. Routledge. pp. 29-58

Appendix 2 Studies accessed in review of the literature

	Location	Sample size: teachers	Sample size: schools	Essential feature 1: Identify research focus	Essential feature 2: Collaborative planning	Essential feature 3: Research lesson with live observation	Essential feature 4: Post-lesson discussion	Supplementary feature a: Linking to what is known	Supplementary feature b: New lesson, no reteach
Black, P.J. (2010)	USA	14	3	Yes	Yes	Yes	Yes	No	No
Grove, M. (2011)	USA	4	1	Yes	Yes	Yes	Yes	No	Yes
Hixon, M. (2009)	USA	8	1	Yes	Yes	Yes	Yes	No	No
Hunter, J. Back, J. (2011)	England	19	4	Yes	Yes	Yes	Yes	No	No
Kriewaldt, J. (2012)	Australia	10	4	Yes	Yes	Yes	Yes	No	No
Lawrence, C. & Chong, W. (2010)	Singapore	10	1	Yes	Yes	Yes	Yes	No	No
Lee, A.T. (2012)	USA	6	1	Yes	Yes	Yes	Yes	No	No
O'Connor, B. (2010)	USA	6	2	Yes	Yes	Yes	Yes	No	No
Podhorsky, C. and Fisher, D. (2007)	USA	30	1	Yes	Yes	Yes	Yes	No	Yes
Puchner, L.D. and Taylor, A.R. (2006)	USA	17	?	Yes	Yes	Yes	Yes	No	No
Yamnitzky, G. (2010)	USA	83	?	Yes	Yes	Yes	Yes	No	No
Cajkler, W., Wood, P., Norton, J. & Pedder, D. (2014)	England	4	1	Yes	Yes	Yes	Yes	No	Yes
Burghes, D. & Robinson, D. (2009)	England	?	?	Yes	Yes	Yes	Yes	No	Yes
McQuitty, V. (2011)	USA	19	2	Yes	Yes	Yes	Yes	Yes	Yes
Yarema, C.H. (2010)	USA	31	?	Yes	Yes	Yes	Yes	Yes	No
Buono, A.G. (2012)	USA	8	1	Yes	Yes	Yes	Yes	Yes	No

Cheng, L.P. and Lee, P.Y. (2011)	Singapore	6	1	Yes	Yes	Yes	Yes	Yes	No
Fernandez, C. (2005)	USA	4	1	Yes	Yes	Yes	Yes	Yes	Yes
Gutierez, S. (2016)	Phillippines	30	?	Yes	Yes	Yes	Yes	Yes	No
Hart, L. (2009)	USA	8	6	Yes	Yes	Yes	Yes	Yes	No
Lewis, C., Perry, R. & Hurd, J. (2009)	USA	6	5	Yes	Yes	Yes	Yes	Yes	Yes
Lewis, C., Perry, R., Hurd, J. & O'Connell, P. (2006)	USA	22	1	Yes	Yes	Yes	Yes	Yes	Yes
Moss, J., Hawes, Z. & Naqvi, S (2015)	Canada	8	1	Yes	Yes	Yes	Yes	Yes	Yes
Moss, J., Hawes, Z., Naqvi, S. and Caswell, B. (2015)	Canada	8	?	Yes	Yes	Yes	Yes	Yes	No
Mutch-Jones, K., Puttick, G. & Minner, D. (2012)	USA	32	?	Yes	Yes	Yes	Yes	Yes	Yes
Norwich, B. & Ylonen, A. (2015)	England	?	?	Yes	Yes	Yes	Yes	Yes	Yes
Roberts, M. (2010)	USA	6	1	Yes	Yes	Yes	Yes	Yes	No
Rock, T. & Wilson, C. (2005)	USA	6	1	Yes	Yes	Yes	Yes	Yes	No
Saito, E., Harun, I. Kuboki, I. & Tachibana, H. (2006)	Indonesia	13	?	Yes	Yes	Yes	Yes	Yes	Yes
Suh, J. & Seshaiyer, P. (2014)	USA	6	1	Yes	Yes	Yes	Yes	Yes	Yes
Verhoef, N., Coenders, F. Pieters, J. van Smaalen, D. & Tall, D. (2015)	Holland	7	4	Yes	Yes	Yes	Yes	Yes	Yes
Wake, G., Foster, M. & Swann, M. (2013)	England	24	9	Yes	Yes	Yes	Yes	Yes	Yes
Ylonen, A. & Norwich, B. (2015)	England	61	?	Yes	Yes	Yes	Yes	Yes	Yes
Cajkler, W., Wood, P., Norton, J., Pedder, D. & Xu, H. (2015)	England	7	1	Yes	Yes	Yes	Yes	Yes	Yes
Droese, S. (2010)	USA	3	3	Yes	Yes	Yes	Yes	Yes	Yes
Fernandez, C. & Yoshida, M. (2012)	Japan	21	1	Yes	Yes	Yes	Yes	Yes	Yes
	Totals:			36	36	36	36	23	19
	Averages:	15.2	1.6						
***Our data (LCK and Ascend combined)**	England	133	33	Yes	Yes	Yes	Yes	Yes	Yes