Ofsted's research into inspector reliability

Alan Passingham

6 March 2020

How we inspect

The role of Ofsted

- The inspection pillar of school accountability
- An inspectorate, not an improvement agency
- Schools inspected on average on a five-year cycle, with a risk-assessed element (providers at risk of reduced quality or failure identified more regularly for inspection)
- Four main criteria evaluated and an overall effectiveness judgement provided
- Four point grading scale: outstanding, good, requires improvement, inadequate. 86% are good or better
- Government takes action on the basis of Ofsted grades

What do we inspect and regulate in education?

- Schools
- Further Education and Skills
- Early years provision
- We are also the regulator of childcare in England
- In addition to our role in education we also inspect and regulate children's social care

Our inspectors

- Two types of inspectors
  - Her Majesty's Inspectors. Full-time employed by Ofsted (approx. 300)
  - Ofsted Inspectors: seconded from education providers (approx. 1700)
- In addition we have regulatory inspectors in Early Years
- All our inspectors are professionals who have worked as leaders in the sectors they inspect or regulate
- They are generalists, but receive regular high-quality training on all areas relevant to their work

Our inspections

- We have two types of school and further education and skills inspections:
  - Section 8 (short inspections)
  - Section 5 (full inspections)

- Outstanding providers are currently exempt from inspection unless risk assessment dictates otherwise.

Our new Education Inspection Framework (EIF)

- Introduced in September 2019
- Response to unintended consequences of the previous framework, which was strongly focused on pupil attainment data
  - This could divert providers from the real substance of education.
  - What young people learn is too often coming second to delivering performance data.
  - This data focus leads to unnecessary workload for teachers.
  - Teaching to the test and narrowing of the curriculum have the greatest negative effect on the most disadvantaged and the least-able children.

Doing research – the *content* of the framework

- Range of research on methodology (e.g. lesson observation)
- Large-scale research programme on curriculum (validity)
- Key factors:
  - The curriculum must be ambitious for all pupils/learners
  - Subject leaders must having clear roles and good subject knowledge
  - Effective curriculum planning is central
  - The curriculum must have sufficient depth and coverage of knowledge
  - A thought out model of progression and sequencing through each subject's curriculum lies at the heart of the curriculum
- This research fed into the development of the criteria in the framework
- It also contributed to the development of a new inspection methodology – 'subject/aspect deep dives'

What is involved in a deep dive?

- Discussion on the intent of the curriculum with senior leaders (what do you want your pupils to know, understand and do)
- Followed by activities that look at curriculum implementation
  - Discussion with subject leaders, teachers and pupils
  - Focus groups or individually
  - Lesson observation
  - Work scrutiny (in isolation or with subject leads)
  - Review of planning documents (scheme of work, etc.)
- Triangulation of the evidence collected results in a quality of education judgement

Reliability study – primary school short inspections

Have we got the right balance?

- Reliability of short inspections (2017) – focus too broad?
  - Looked at all aspects of the framework
  - Completed on live inspection (which was the priority)
  - Lots of triangulation points
- Lesson observation (2019) – focus to narrow?
  - Judgements determined solely from observation
  - But identified that triangulation with other evidence sources is required
- FES reliability study (2020) – the right balance?
  - Quality of education judgement only (not the whole framework)
  - Not live inspections (research findings the priority)
  - Focus on the new deep dive methodology

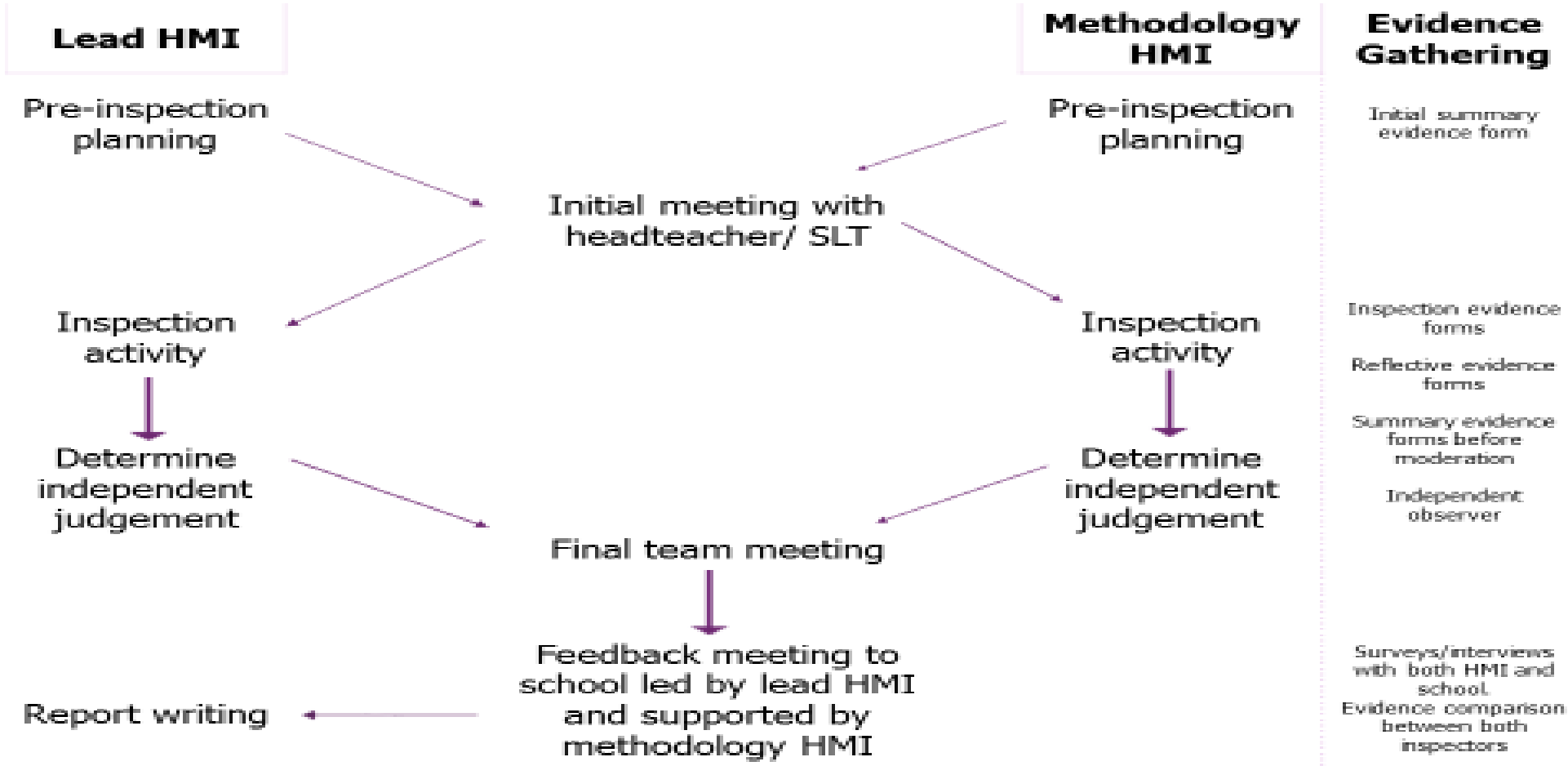
Key questions

- Are inspection judgements reliable for short inspections?
- Is the reliability testing method an effective approach for establishing reliability in short inspections?

Methodology

- Two inspectors inspecting the same school at the same time
- Carried out on live inspection
- Complex design –
  - Inspection outcome remained the priority
  - Artefacts built in to minimise inspection burden on the provider (i.e. joint discussions led by one of the inspectors)
  - Clarity of 'lead' and 'methodology' inspector roles (inspector shadowing)
  - Arrangements needed for converting a short inspection to a full inspection
  - Managing the independence of the two inspectors (data spill-over)
  - Use of reflective evidence forms at points during the visit
- Scored using a three-point scale

Process


Lead HMI
Methodology HMI
Evidence Gathering
Pre-inspection planning
Pre-inspection planning
Initial summary evidence form
Initial meeting with headteacher/ SLT
Inspection activity
Inspection activity
Inspection evidence forms
Reflective evidence forms
Summary evidence forms before moderation
Independent observer
Determine independent judgement
Determine independent judgement
Final team meeting
Feedback meeting to school led by lead HMI and supported by methodology HMI
Surveys/interviews with both HMI and school
Evidence comparison between both inspectors
Report writing

Sample size was smaller than hoped

- Intention to sample 80 schools (based on estimate of inter-rater agreement at 0.8 and confidence interval +/- 10%)
- 54 inspections arranged
- Total of 26 inspections carried out
- Results of 24 valid

    - one inspection, inspector judgements were not formed independently

    - one inspection converted due to safeguarding issue

Short inspections appear reliable…

- In 22 of the 24 completed inspections, inspectors arrived at the same judgement (92%).
- Where inspectors interpreted the evidence or where different activities were conducted (or indeed the same activities but in a different order), inspectors nevertheless reached the same view of the school overall.
- Two inspections where there was a difference in outcomes
  - Inspector subjectivity
  - Different pathways through the inspection that missed key elements

…but some design issues apparent

- Small sample size
- Not a true parallel inspection
- Resource implications
- More burdensome process for schools with weaknesses

How valid and reliable is lesson observation?

International seminar on lesson observation

Overarching questions:

- What can Ofsted learn from international best practice in the use of classroom observation for the evaluation and improvement of teaching quality?
- What changes should Ofsted consider in revising its inspection framework?

The six models

- The Classroom Assessment Scoring System (CLASS)
- Framework for Teaching (FfT)
- The International Comparative Analysis of Learning and Teaching (ICALT)
- The International System for Teacher Observation and Feedback (ISTOF)
- The Mathematical Quality of Instruction (MQI)
- Generic Dimensions of Teaching Quality

Main findings from the seminar

- Validity just as important as reliability.
- The six models all used scaled indicators but, the experts were clear that these were still high inference models. Primacy was given to expert judgement.
- Indicators were a means to provide a valid structure to observation.
- Deciding the purpose of the model is essential. The experts agreed that in our context there was no rationale to focus on teacher or lesson level observation. Aggregation of observations at the school level was considered much more useful.
- Learning is invisible. Instead, a focus on teachers and teaching ensures validity in an observation model.
- Observation alone does not explain everything about the quality of teaching; other sources of evidence are also required.
- A high standard of training and regular refresher trainer in using an observation instrument are essential for reliability.

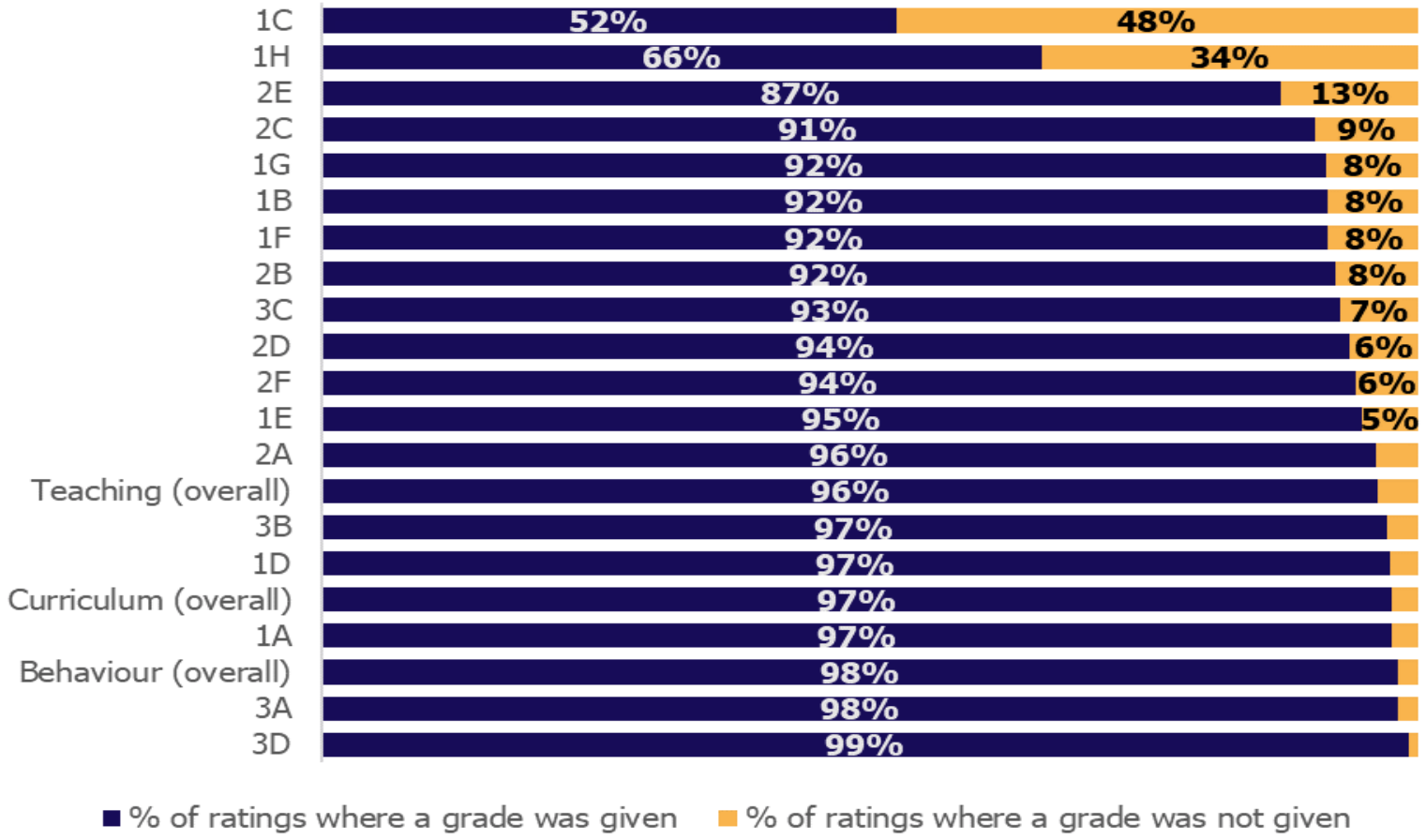
The seminar informed our research

- Focus:
  - How valid and reliable is the use of lesson observation in supporting judgements on the quality of education?
- Objectives
  - Test a series of indicators that could underpin lesson observation practice in the new framework
  - Identify important practices and structures that elicit valid and reliable evidence to support judgements on the quality of education
  - Test whether observation can be effectively assessed at the department level for evaluation purposes.

Method applied

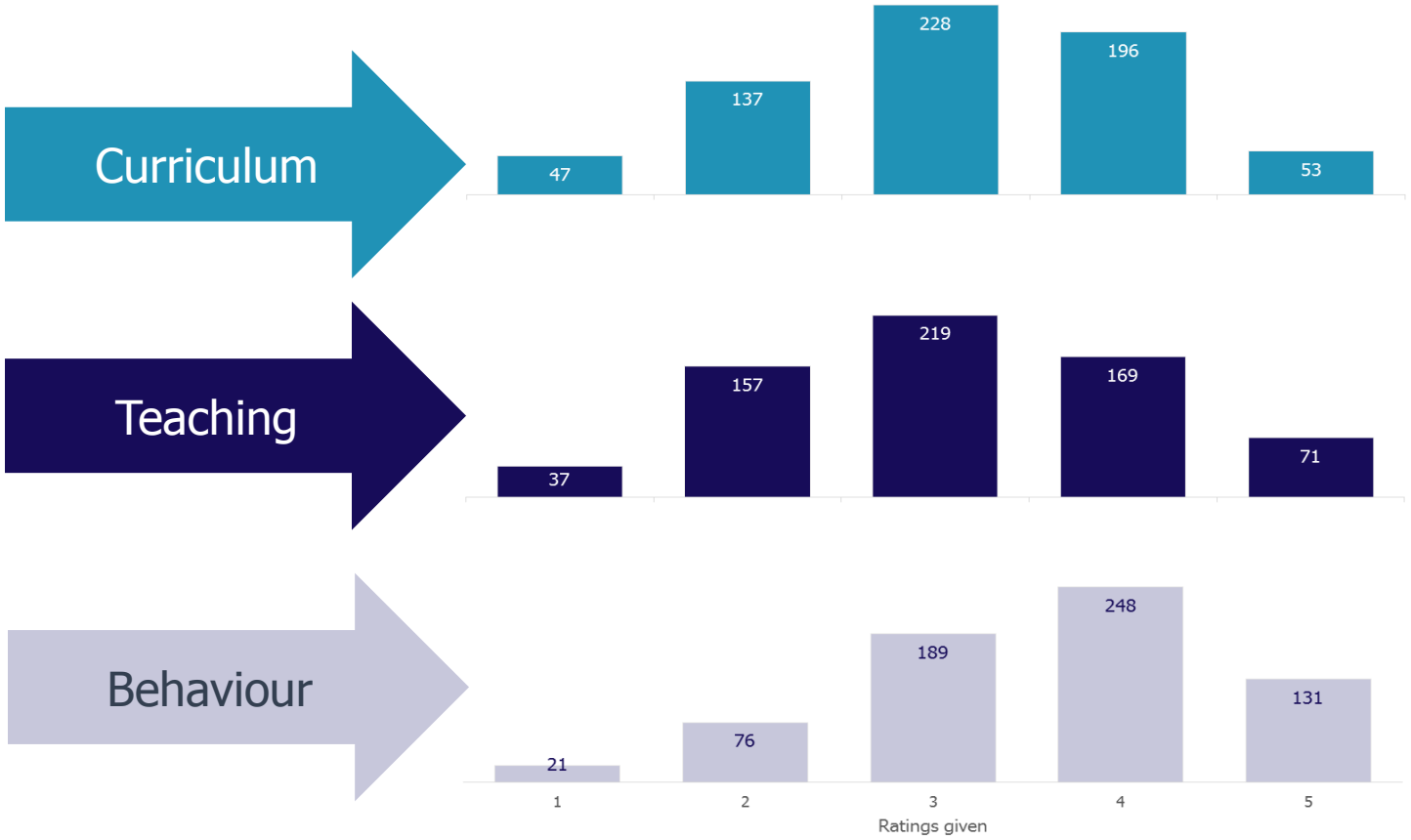
- Paired inspectors observing the same lesson – nine HMI and four researchers participated in study
- Observation instrument designed to assess 18 indicators
  - Spread over three domains: **curriculum**, **teaching** and **behaviour**
- Indicators developed with our inspection purpose in mind (how to assess the quality of education)
- Observations grouped by 'department' (two departments per visit)
- Judgements made on a five-point scale at individual lesson and department level (detailed rubric included to provide structure and aid consistency)
- 37 providers visteid (22 schools, 15 colleges) with 346 lessons observed across 63 departments
  - Two HMI focus groups provided further evidence

Indicators for reading/numeracy strategies and assessment were difficult to observe



This suggests these are less useful indicators for inspectors to focus their attention on during observation.

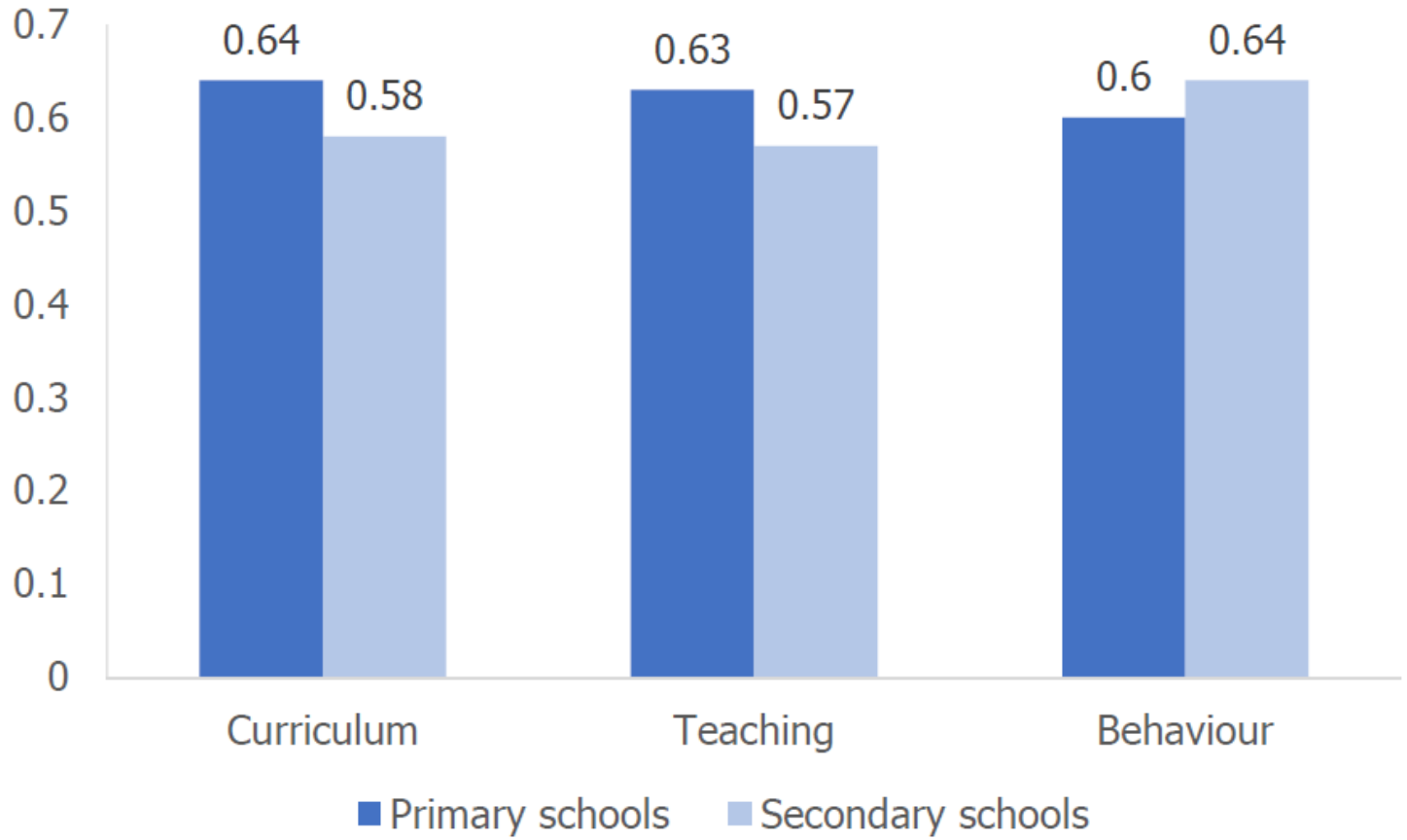
Behaviour indicators were scored more strongly than those for teaching and curriculum



This provides our model with face validity as this pattern replicates that found in other lesson observation models.

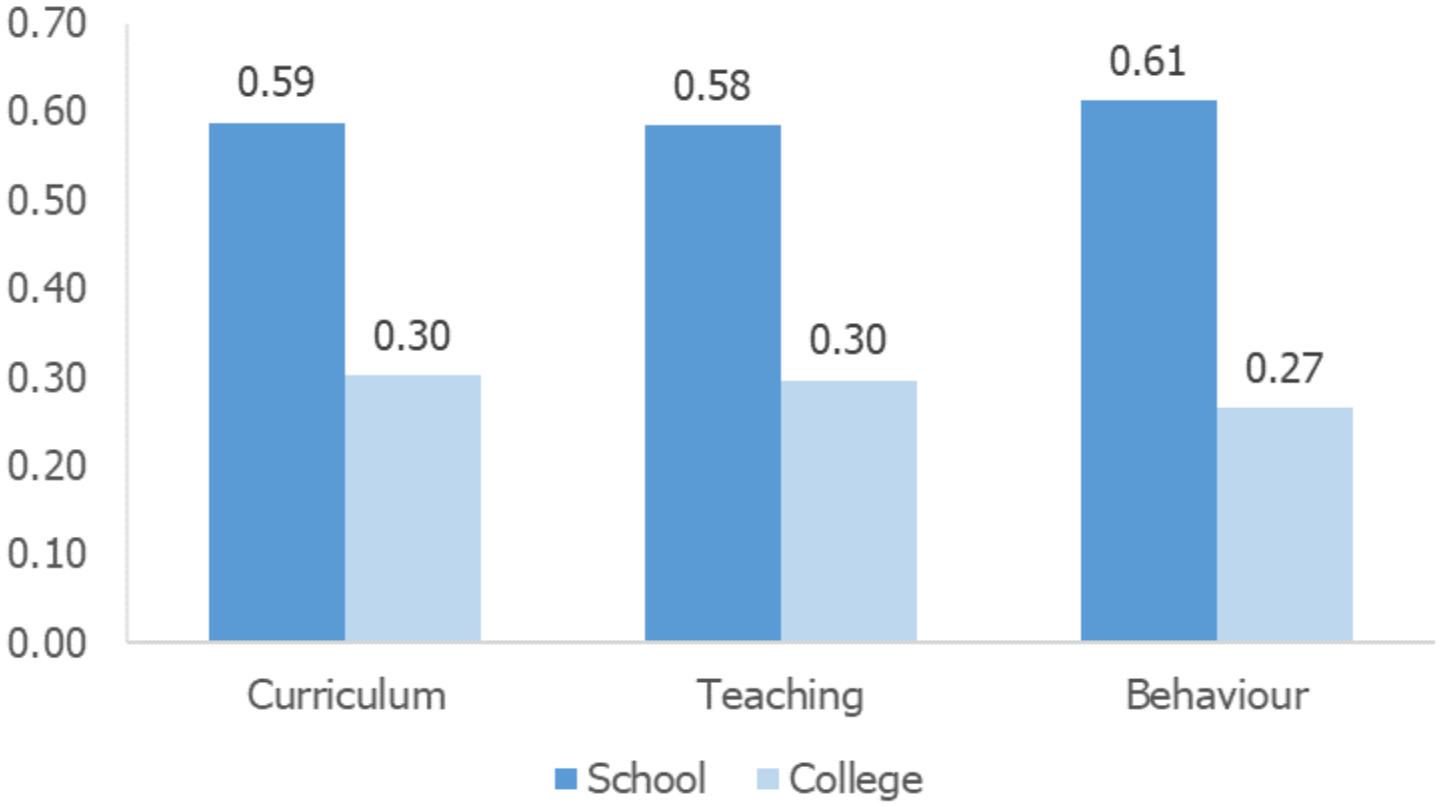
It suggests that while pupils are compliant in most lessons, this does not mean they are always being taught well.

Primary school observations were reaching a substantial level of reliability. Reliability in secondary schools was slightly lower.

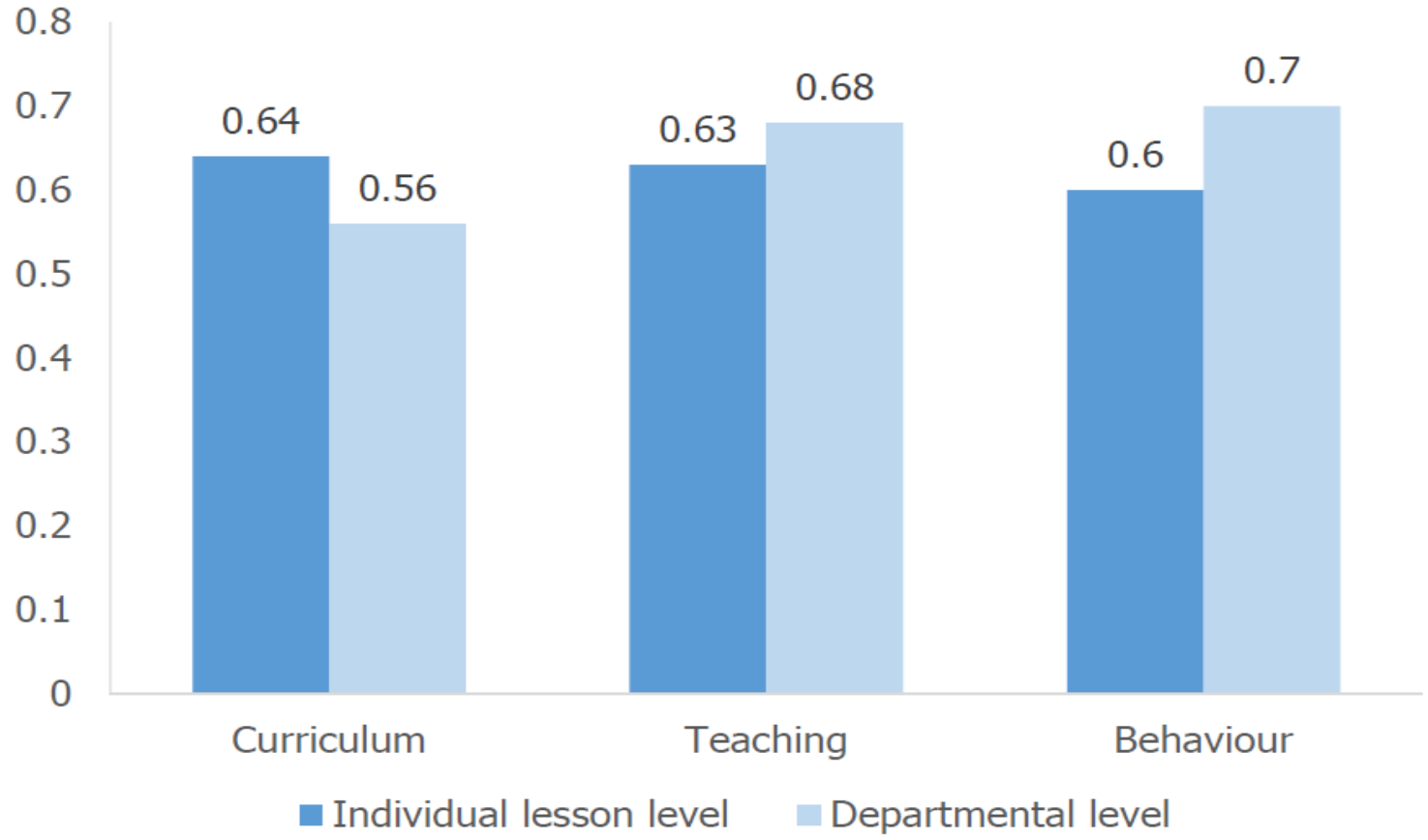


Kappa statistic	Agreement
0<x≤0.2	Slight
0.2<x≤0.4	Fair
0.4<x≤0.6	Moderate
0.6<x≤0.8	Substantial
0.8<x≤1	Almost perfect

Reliability in colleges, however, only reached a fair level of reliability



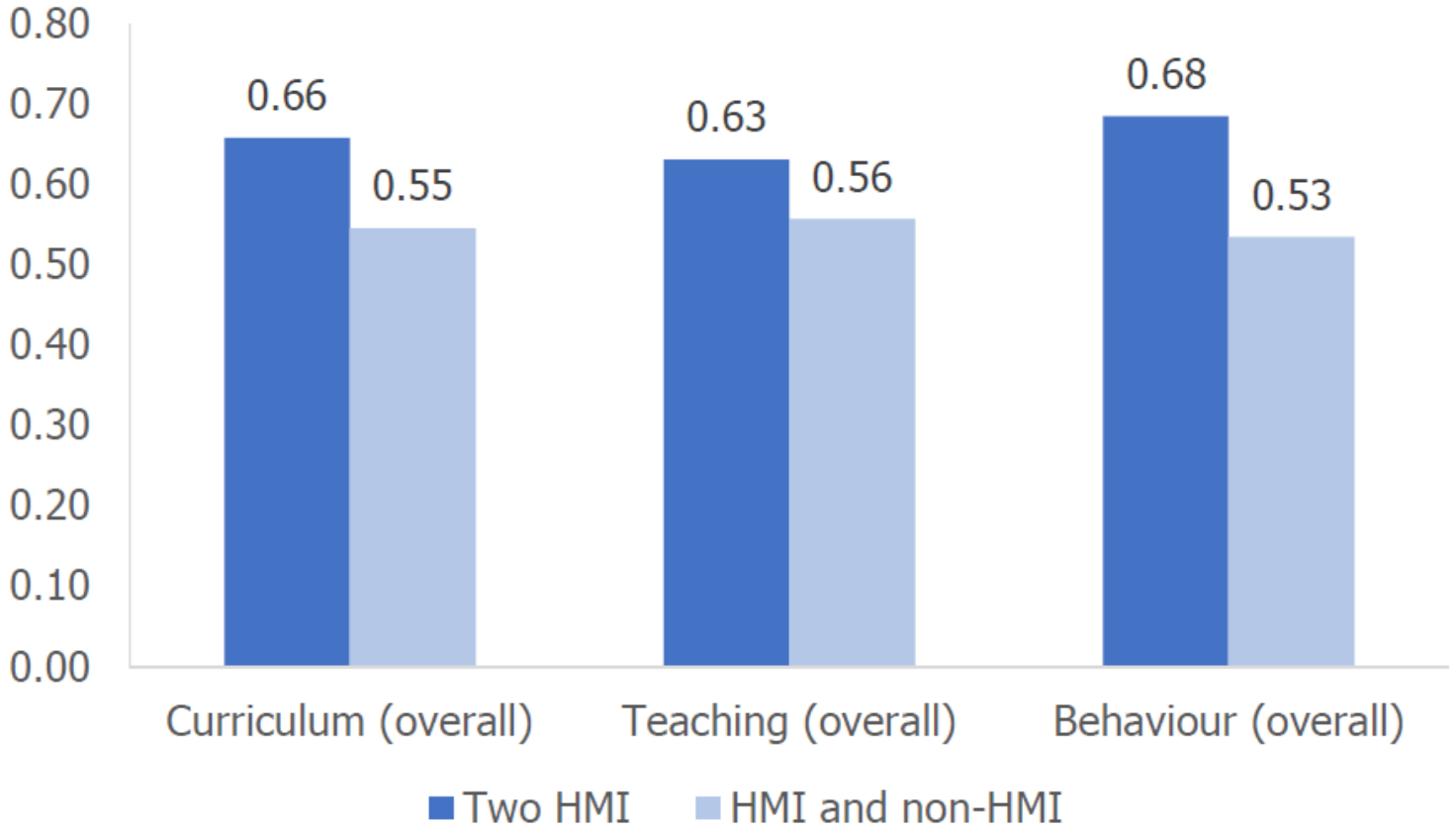
For primary schools, strong reliability found when lesson scores were synthesised at the department level



Department scores in secondary schools were less reliable though (moderate reliability achieved in curriculum and teaching domains).

This was related to inspectors carrying out observation in subjects were they lack expertise.

Overall scores generally improved with inspector expertise



Other analyses identify that…

- Observation length (15 to 30 minutes) appears to make little difference to reliability
- Reliability increased in the departments visited in the afternoon suggesting a practise effect may be enhancing the level of consistency.

The level of reliability achieved, particularly in colleges, could be explained by several factors

- Lack of standardised training for inspectors
- Cognitive load on inspectors
- Differing contexts (schools and colleges)
- Study not true to live inspection
- Issues with central tendency
- Subject expertise (in secondary schools)

Implications for our new framework

- Indicators and structure of our model is valid.
- Reducing the number of indicators should enhance consistency.
- Training programmes need developing to improve inspector reliability – structuring observation only gets us part way there.
- Triangulation required - reliability data suggests observation is just one part of the picture that explains quality of education.
- Where appropriate multiple inspectors should be present for all inspections as collaboration is likely to improve consistency.
- More research is required.

Note

- The principles behind the lesson observation model are used on inspection, but the scoring mechanism is not.
- Scoring is being used, however, in our evaluation work on the Education Inspection Framework.

College reliability study

Research questions

- Focus
  - Do college inspectors reach reliable judgements about the quality of education when using the deep dive method of inspection?
  - What makes it reliable?
  - What hinders reliability?

- Objectives
  - To test inspector reliability in judging the quality of education through the deep dive methodology in college providers.

Approaches to testing reliability

- Two stage approach which will provide qualitative and quantitative evidence

1) **fieldwork in a 'live' capacity** through research visits focused on the deep dive process rather than live inspections.	2) **a desk-based approach** where inspectors make comparable judgements on the basis of a review of evidence from already completed college inspections.

Method

Similar design to our original reliability study, but focused on two teams of inspectors and just the deep dive process.

Pros

- **Removal of live inspection component** and a focus on a single aspect of the inspection process (the deep dive) should mitigate against sample attrition.
- Captures **collaboration** between inspectors that is often found in judgement formation on a FES inspection.
- Along with a single indicator – quality of education on a four-point scale – this increases **construct validity**.

Cons

- Likely to **reduce sample size**, owing to more of the available inspector resource being used up to team on visits.
- Large inspection teams will be a burden for providers.

Main design features

- Two teams of inspectors to test reliability
- This will comprise four inspectors per team
- **Research visit** spread over three days to replicate, as far as possible, what happens in the deep dive process and judgement formation on inspection
- Medium-sized or larger providers to accommodate capacity of both teams
- Inspectors decide on subject focus and activities to use during the deep dive (no process derived by research team)

Reliability design

1. **Collaboration**:
   - Inspectors will be deployed in two teams of four
   - They will engage in consultative team-based approach to collect evidence and reach judgements
2. **Managing independence**:
   - Each team accompanied by a neutral observer.
   - They will ensure minimal contact between the two teams.
   - This should maintain the independence of the QE judgement that each team provides.

Implications for 'live' research visits

- No 'artefacts' being introduced i.e. additional quantitative scoring
- There may be a practice effect but this still improves on a 'shadowing' approach
- There may be a burden on the provider though the selection of the same subjects for the deep dive
- Need to be aware of spillage from external sources, such as nominees, other provider staff, etc.
- A degree of inspector blinding required – involvement of neutral observers, separate hotels, etc.

Main issues

- Bigger inspection teams = reduced sample size.
- Sample attrition. These are research visits – not inspection – so consent is required from provider leaders.
- Only four visits possible with available spring term resource.
- More visits expected in the summer term, but very unlikely to reach a statistically generalisable scale.
- Phase 2 approach required – desk-based reliability approach will be built into the back-end of the study.
- This may tell us something about the reliability of our quality assurance processes.

What does the evidence currently tell us?

Some drivers for reliability suggested

- Structure (ensuring that the framework and handbook are valid)
- Triangulation of evidence from different activities essential
- Inspector collaboration
- Inspector training
- Post-visit quality assurance procedures

However

- There remains a small trade off between validity and reliability
- 'Perfect reliability' difficult to attain for complex, real-world judgements
- Subjective judgements vs test score validity

Resources

- Do two inspectors inspecting the same school make consistent decisions?
- Six models of lesson observation: an international perspective
- Curriculum research: assessing intent, implementation and impact
- Inspecting education quality: lesson observation and workbook scrutiny
- The nuance of reliability studies

Ofsted on the web and on social media

www.gov.uk/ofsted

https://reports.ofsted.gov.uk

www.linkedin.com/company/ofsted

www.youtube.com/ofstednews

www.slideshare.net/ofstednews

www.twitter.com/ofstednews