Working Paper No. 23-03

Are some school inspectors more lenient than others?

Christian Bokhove
University of Southampton

John Jerrim
University College London

Sam Sims
University College London

School inspections are a common feature of education systems across the word. These involve trained professionals visiting schools and reaching a high-stakes judgement about the quality of education they provide. By their nature, school inspections rely upon professional judgement, with different inspectors potentially putting more emphasis on certain areas than others. Yet there is currently little academic evidence investigating the consistency of school inspections, including how judgements vary across inspectors with different characteristics. We present new empirical evidence on this matter, drawing upon data from more than 30,000 school inspections conducted in England between 2011 and 2019. Male inspectors are found to award slightly more lenient judgements to primary schools than their female counterparts, while permanent Ofsted employees (Her Majesty's Inspectors) are found to be harsher than those who inspect schools on a freelance basis (Ofsted Inspectors).

VERSION: February 2023

Suggested citation: Bokhove, C., Jerrim, J., & Sims, S. (2023). *Are some school inspectors more lenient than others?* (CEPEO Working Paper No. 23-03). Centre for Education Policy and Equalising Opportunities, UCL. https://EconPapers.repec.org/RePEc:ucl:cepeow:23-03.

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the IOE, UCL's Faculty of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- Male lead inspectors awarded a low grade ('Requires Improvement' or 'Inadequate') to around a third (33.1%) of primary schools. By contrast, female lead inspectors awarded a low grade to 36.4% of primary schools.
- This difference was particularly pronounced for the lowest inspection grade – 'Inadequate' – which can often result in headteachers of the inspected schools losing their jobs. Female lead inspectors were one third more likely to award an 'Inadequate' grade to primary schools than their male counterparts (5.9% versus 4.5% respectively).
- These patterns remained when comparing male and female inspectors sent to inspect schools with the same prior Ofsted inspection rating, exam results, levels of pupil absences, pupil intake, and in the same region of the country.

Why does this matter?

'Inadequate' Ofsted judgements can lead to schools closing and headteachers losing their jobs. High-stakes assessments of this sort should be made in a reliable way.

Are some school inspectors more lenient than others?

Christian Bokhove (University of Southampton)
John Jerrim (UCL Social Research Institute)
Sam Sims (UCL Centre for Education Policy and Equalising Opportunities)

Author list alphabetical. All joint first authors.

February 2023

School inspections are a common feature of education systems across the word. These involve trained professionals visiting schools and reaching a high-stakes judgement about the quality of education they provide. By their nature, school inspections rely upon professional judgement, with different inspectors potentially putting more emphasis on certain areas than others. Yet there is currently little academic evidence investigating the consistency of school inspections, including how judgements vary across inspectors with different characteristics. We present new empirical evidence on this matter, drawing upon data from more than 30,000 school inspections conducted in England between 2011 and 2019. Male inspectors are found to award slightly more lenient judgements to primary schools than their female counterparts, while permanent Ofsted employees (Her Majesty's Inspectors) are found to be harsher than those who inspect schools on a freelance basis (Ofsted Inspectors).

Key Words: Ofsted, school inspection, consistency, accountability.

Contact details: Christian Bokhove, University of Southampton. C.Bokhove@soton.ac.uk

https://doi.org/10.5258/SOTON/P1108

Acknowledgements: The Nuffield Foundation is an independent charitable trust with a mission to advance social wellbeing. It funds research that informs social policy, primarily in Education, Welfare, and Justice. It also funds student programmes that provide opportunities for young people to develop skills in quantitative and scientific methods. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics and the Ada Lovelace Institute. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org. We are grateful for their support. Helpful comments have been received on the draft from our project advisory group, whom we would like to thank for their input and support.

1. Introduction

School inspections are an important feature of many education systems. These involve a team of trained inspectors visiting schools and judging the quality of education that they provide. The outcomes of these inspections are often high stakes for schools and their staff (Kemethofer, Gustafsson, & Altrichter, 2017). Judgements are often made publicly available, get widely reported by local media and can – in the extreme – lead to school closures or the removal of headteachers (Eyles & Machin, 2019). Data and reports from inspections also get widely used by a variety of key stakeholders, including parents when they are choosing schools (Ofsted, 2019c). Thus, given the importance attached to inspection outcomes, it is vital they are as valid, consistent and reliable as possible. Inspectorates – such as Ofsted in England – thus devote significant time and resource into developing inspection frameworks, and then training inspectors in their use (Ofsted 2019d).

Yet, despite these efforts, some have questioned the validity of Ofsted inspections (whether they accurately capture school quality) and the consistency of outcomes across different inspectors (whether the same judgements would be made if the inspection were conducted by different inspectors or on different days). Despite the significant effort inspectorates put into developing and providing training in their frameworks, evidence from the public administration literature has questioned how much control central government bodies have over the actions of their front-line employees (Ingersoll, 1993). Moreover, the subjective nature of inspection means that a degree of human judgement will always be involved (Spielman, 2017), with some inspectors putting greater emphasis on certain aspects of schools than others. This is also recognised within inspection handbooks, which note how inspectors should "*draw on all the evidence they have gathered and using their professional judgement*" (Ofsted, 2022a). Yet this has led some in the education sector to question the usefulness of school inspections as a mechanism for monitoring school standards and as a force for improvement (National Education Union, 2021). There is particular concern that inspection outcomes may be influenced – at least in part – by factors outside of a school's control (Richmond, 2019). This includes, for instance, the characteristics of the inspector(s) they happened to be assigned.

A relatively small number of empirical studies have previously been conducted into the consistency of school inspections, including in England (the empirical setting of this paper). Moreover, much of the existing work has been conducted by school inspectorates themselves.

Around 25 years ago, a research project led by Ofsted investigated whether different inspectors observing the same lesson awarded it the same grade (Matthews et al., 1998). Collecting data from 100 inspections, encompassing 173 pairs of inspectors, they found the correlation between the judgements made by different inspectors to be high (Pearson correlation = 0.81; Cohen's Kappa = 0.53). They hence conclude that their results "*suggest that OFSTED's framework and related advice provide an effective means by which such inspectors can judge teaching with considerable reliability*". More recently, Ofsted (2017) investigated the consistency of its short inspections based upon a sample of 24 schools (all of which were previously judged to be Good). Specifically, each school was assigned two inspectors, with their judgements (whether to convert the inspection or not) then compared. In 22 of the 24 schools the same inspection outcome was reached. This work, and the findings it presents, have however come under some criticism (Pearson, 2018).

Other research published by Ofsted has focused upon the inter-rater reliability of specific inspection tasks. One example is where nine of Her Majesty's Inspectors (HMIs) undertook "workbook scrutiny", where the same documents were evaluated by two or three independent inspectors (Ofsted 2019a). This found there to be "moderate" levels of inter-rater reliability across four indicators (building on previous learning, depth and breadth of coverage, pupils' progress and practice), with Cohen's Kappa standing around 0.5[1]. The English school inspectorate also conducted a similar investigation into the reliability of lesson observations (Ofsted, 2019b). This reported moderate-to-substantial levels of reliability (Kappa statistics around 0.6) for schools, though much lower levels for colleges (Kappa statistics around 0.3). They also found greater levels of consistency between two of Her Majesty's Inspectors (HMIs) than between an HMI and a (freelance) "Ofsted inspector".

Despite the valuable insights gained from these studies, there remains some notable limitations with the existing evidence base. The important work previously published by Ofsted is based on a small number of inspectors, with it not being clear the extent that the results can be generalised across the inspection workforce. In particular, much existing work surrounding inspection consistency has mainly involved Her Majesty's Inspectors (HMIs), who are full-time Ofsted employees. Yet, in the real-world, many Ofsted inspections are conducted by freelancers (Ofsted Inspectors), without any HMI involvement (further details are provided in

[1] To put this figure into context, Jerrim and Micklewright (2014) report similar levels of consistency (based upon Cohen's Kappa) for how reliably 15-year-olds report their parents' level of education.

the following section). Moreover, those HMIs that participated in the previous investigations conducted by Ofsted would have been aware they were involved in a research study, which may have impacted upon their behaviour. To draw an analogy, if an examination marker knew that someone else was going to mark the same paper – and that this was going to be used to make judgements about the reliability of their marking – then they are likely to complete the task as conscientiously as possible (potentially more so than usual). The same may hold true with respect to the behaviour of inspectors in such research studies. In contrast, little work has focused upon variation in inspection outcomes across different inspectors when they are conducted in a more "natural" setting (i.e. using data from real-life inspections). Likewise, more generally, there is little existing work exploring the extent that inspection outcomes vary depending upon the inspector(s) that schools are assigned. Finally, much work has been conducted by inspectorates themselves, rather than by independent academic groups.

This paper aims to start to fill these gaps in the literature. Using data from more than 30,000 school inspections conducted in England between 2011 and 2019, we present novel evidence on how inspection outcomes vary across different (lead) inspectors, including how this is related to their observable characteristics. Specifically, we estimate the proportion of the variation in inspection outcomes that occurs across different inspectors, and whether the judgements reached depend upon factors such as the lead inspector's gender, contract type and experience. In doing so, we provide (to our knowledge) the first piece of independent research into such "inspector effects", either in England or worldwide. More generally, the paper presents one of the first pieces of independent academic evidence on the topic of inspection consistency.

2. Background to school inspections in England

What are Ofsted inspections?

Since the early 1990s, school inspections in England have been conducted by the Office for Standards in Education (Ofsted) – a non-ministerial government department (https://www.gov.uk/government/organisations/ofsted). Prior to inspection, inspectors are provided with a range of background information about the school, including details about its pupils, previous inspection rating, absence levels and recent performance in national

examinations. During an inspection, a team[2] of inspectors visit the school, where they will observe lessons and gather evidence about the quality of education it provides. They will also seek the views of parents, pupils and teachers (through both surveys and conversations) while liaising with senior school staff. Based upon their assessment of the evidence collected, the inspectors will use their professional judgement to reach a verdict about the school's overall quality (see below for further details). They will also produce a summary of the evidence in a written report, made publicly available on the Ofsted website (https://reports.ofsted.gov.uk/).

Full versus short inspections

Most Ofsted inspections fall into one of two broad categories. The first is "full" (section 5) inspections. These are longer, more in-depth inspections that typically last for around two full days (Ofsted, 2022b). Full inspections lead to schools receiving one of four Overall Effectiveness judgements:

1. Outstanding
2. Good
3. Requires Improvement
4. Inadequate

Schools also receive one of these four grades in various sub-domains, such as "*leadership and management*", "*outcomes for pupils*" and "*quality of teaching, learning and assessment*". While the four-point Overall Effectiveness judgement has been used consistently since 2011 (when our study period starts), the sub-domains schools are judged upon has changed over time (see below for further details).

The second main inspection type is "short inspections". These were introduced in September 2015 and reserved for schools judged to be Good or Outstanding during their last inspection. Between September 2015 and August 2019 these inspections typically lasted for a single day, although this increased from September 2019 to two days for schools with more than 150 pupils. Short inspections tend to be lighter touch, and do not require the lead inspector to award an Overall Effectiveness grade. Rather, they can either:

(a) Confirm that the school remains at its previous grade (e.g. a school previously judged as Good remains Good).

[2] This team may vary in size from a single individual to five inspectors or more, depending upon the size of the school and the nature of the inspection taking place.

(b) Should convert to a full inspection "immediately" (within 48 hours up until January 2018, changed to within a maximum of seven days after January 2018).
(c) Recommend that a full S5 inspection should be conducted within the next one to two years. This option was introduced in January 2018, with an inspector using it if they felt that the school was either at risk of declining to "Requires Improvement" or improving to "Outstanding".

Schools judged to require improvement or to be Inadequate may also be subject to other forms of inspection, such as re-inspections or monitoring visits. Ofsted may also conduct no-notice inspections where they have concerns (e.g. because of their risk assessment or receiving complaints). The distribution of all types of inspections conducted between September 2011 and August 2019 by academic year can be found in Appendix A.

The frequency and consequences of Ofsted inspections

All new schools are typically inspected within their first three years. For existing schools, the date of the next inspection depends how it was judged during its last inspection. An overview is provided in Table 1, along with the "consequences" of receiving different outcomes. In summary, schools that receive more negative ratings are (a) inspected more frequently and (b) more likely to receive a full rather than a short inspection next. For schools deemed to be Inadequate, they are either served with an academy order (i.e. a forced change of management) or be subject to monitoring visits. It is also likely to lead to a change of headteacher (Eyles & Machin, 2019) and is thus a particularly high-stakes judgement for inspectors to make.

<< Table 1 >>

Who are Ofsted inspectors?

Ofsted employ two categories of inspector. The first are Her Majesty's Inspectors (HMIs). These individuals are permanent, usually full-time Ofsted employees. They are employed as civil servants and work for Ofsted as their only job. The second category are Ofsted Inspectors (OIs). These individuals work for Ofsted as freelancers and conduct inspections on an ad-hoc basis. OIs typically hold other jobs, with many being education professionals working in schools (e.g. as headteachers or other senior school leaders). Up until September 2015, OIs were employed by private sector organisations such as Serco. They have however since been

directly contracted by Ofsted. This led to a sharp decline in number of OIs - from around 3,000 to 1,600 - with it claimed those released were "not Good enough" (Richardson, 2015).

Key changes to Ofsted inspections between September 2011 and August 2019

Our analysis focuses on inspections conducted between September 2011 and August 2019. There were several changes to Ofsted inspections during (and since) this period which may, at times, have implications for the interpretation of our results. The key points, in chronological order, are as follows:

- 2012. Changes were made from the previous school inspection framework in January and September 2012 (Ofsted, 2012). This included exempting schools previously judged to be Outstanding from routine inspections. See Richards (2012) for further discussion.
- September 2012. Prior to this date, the third rung on Ofsted's four-point judgement scale was labelled "satisfactory". In September 2012 this changed to "Requires Improvement". This to some extent changed expectations, with it made clear to schools that they should not remain at this level.
- January 2013. Ofsted moved to its current regional structure. England was divided into eight regions, each led by a regional director who was responsible for managing and delivering inspections. This replaced more centralised direction of these activities.
- September 2015. Several changes were made to Ofsted inspections. This included the introduction of a common inspection framework across its different remits, the introduction of short inspections, and OIs now directly contracting with Ofsted (rather than through commercial providers). This was also the period in which there was a substantial decline in the number of OIs inspecting on behalf of Ofsted.
- January 2018. Ofsted noted how "*the process for converting short inspections to full section 5 inspections has proven challenging for both schools and inspectors*" (Ofsted, 2017b). This was because conversion required short notice changes to the inspection schedules of OIs, many of whom also hold other jobs. Indeed, Ofsted noted how "*OIs are typically busy school leaders who have booked time off to inspect, and these last minute changes are frustrating and impractical*" (Ofsted, 2017c). Consequently, from January 2018, rather than issuing an immediate conversion, inspectors also had the option of recommending a full inspection to be conducted next (within the next year or

two). As Table 2 illustrates, this coincided with a notable reduction in the number of short inspections being immediately converted.

- September 2019. The new Education Inspection Framework (EIF) was introduced. This led to several significant changes to how schools were inspected, including a greater emphasis on the curriculum and less weight given to school test results.

<< Table 2 >>

As discussed in the sections that follow, our main analysis will pool data from across the September 2011 – August 2019 period to maximise sample size. However, we will also present some estimates separately by academic year or time period.

Research questions

We begin by providing a basic piece of descriptive evidence currently missing from the literature; to what extent do inspection outcomes vary between different lead inspectors? This will in-turn provide a broad, aggregate overview of overall differences in inspection outcomes that lead inspector's reach.

RQ1. What proportion of the variation in inspection outcomes occurs between (rather than within) inspectors?

Our investigations then turn to whether the harshness/leniency of inspectors are related to their background characteristics. We start with gender. A wide body of evidence has found important gender differences in decision making processes (Villanueva-Moya & Expósito, 2021) with it reported that "*men decide faster, more lineal, whereas women gather information in a different way and are more aware of informal sources of information*" (Gernreich & Exner, 2015). Evidence from criminology has also found female judges to impose harsher sentences than males (Steffensmeier & Hebert, 1999). In contrast, male and female assessors were found to provide roughly equal scores to candidates in the context of medical examinations (McManus, Thompson, & Mollon, 2006). Yet there is currently no analogous evidence with respect to gender differences in the judgements made by school inspectors. Our second research question is therefore:

RQ2. Do female inspectors make harsher or more lenient judgements about schools than their male counterparts?

Inspection outcomes may also differ between inspectors with different contract types. Recall that OIs work for Ofsted on a freelance basis, with many working in schools as part of their day job (e.g. as part of school leadership teams). On the other hand, HMIs – as full-time Ofsted employees – do not currently work "at the coal face" leading schools. This may generate differences in the views of HMIs and OIs in what constitutes Good practice, and in their understanding of young people's educational and pastoral needs. OIs may be more "in-touch" with the current challenges facing the teaching profession. Moreover, another key difference from HMIs is that OIs may have recently (or will soon be) subject to Ofsted inspections themselves. Evidence from the management literature also suggests that employees with different contract types may differ in their motivation (Grund & Thommes, 2017), work-related expectations and commitment (Süß & Kleiner, 2007). Such factors may also influence how inspectors go about their job, and thus the judgements that they reach. The third research question therefore asks:

RQ3. Do Ofsted inspection judgements differ between OIs and HMIs?

Next, we turn to the link between inspection outcomes and the lead inspector's experience. Evidence from elsewhere in the education literature (e.g. on teacher effectiveness) illustrates how experience is linked to staff effectiveness and productivity (Burroughs et al., 2019). Moreover, employees new to their roles tend to be less confident, and more liable to make mistakes, than more senior staff (Grohnert, Meuwissen, & Gijselaers, 2019). Indeed, experience in jobs is linked to competence development (Paloniemi, 2006). On the other hand, newly appointed inspectors may be concerned about making potentially controversial, high-stakes decisions when they are fresh into the role (e.g. awarding an Inadequate judgement or downgrading a school). Hence (in)experience could be a key source of inconsistency (and thus variation in outcomes) across inspectors. Our fourth research question is therefore:

RQ4. How are inspection outcomes linked to inspection experience of lead inspectors?

Next, we consider where the inspection is taking place. Ofsted's regional operating model means that inspectors will usually conduct their inspections within one of Ofsted's eight

regions (their "home region"). Although all regions inspect to a common framework, with a certain degree of centralised guidance and training, regions also have autonomy over delivering and managing inspections. It is hence possible that, when an inspector works outside of their home region, they come across certain practises and approaches that they are not use to. Moreover, there may also be regional differences in how schools operate that impact the judgements that inspectors reach. We investigate this in our next research question, providing the first evidence as to whether inspectors award harsher or more lenient judgements when working outside their home region:

RQ5. Do inspectors judge schools more harshly when they are working outside of their home region?

School inspectors will have specialist knowledge, background, and skills in particular areas. One of the most important is whether they have a background in primary or secondary education (and thus primary or secondary inspections). Yet England has many more primary schools than secondary schools, meaning that there are also more primary school inspections that require a lead inspector. This invariably means that some inspectors who have knowledge and inspection experience in one school phase (e.g. secondary) will sometimes lead inspections in another phase (e.g. primary). There are clear ways that this may impact upon the inspection judgement made. For instance, those with a specialism / background in secondary inspections may "play it safe" when asked to inspect a primary school, given that they have less inspection experience in this area. They may thus shy away from issuing potentially high-stakes grades (e.g. Inadequate judgements). Alternatively, secondary schools in England tend to receive lower Ofsted grades than primary schools (e.g. in 2020, 88% of primary schools were rated as Good or Outstanding, compared to 76% of secondary schools; Ofsted, 2020). Inspectors who usually inspect secondary schools may hence also award harsher grades to primary schools. We thus investigate this issue in our sixth research question:

RQ6. Do inspectors with a specialism in secondary school inspections judge primary schools more harshly than inspectors with a primary specialism?

Finally, some school inspections are carried out by a single inspector rather than by a team (Table 3 below provides further details). Yet previous research has noted how, when making decisions, "*individuals are more likely to be influenced by biases, cognitive limitations, and*

social considerations" than groups (Charness & Sutter, 2012). This is potentially due to the benefits of pooling information, the ability to have open discussions about the evidence, or the ability to overcome hidden or unconscious biases and drawing upon the wisdom of groups (Bang & Firth, 2017). Indeed, within the broader literature of inspection, research has found that "*groups of inspectors produced more reliable assessments than individual inspectors*" in the context of hospitals. Yet, in terms of optimal team size, the evidence remains inconclusive – although somewhere in the range of between 5 and 12 team members is often cited (Powell & Lorenz, 2019). Moreover, the potential advantages of larger teams may be dissipated if it leads to "groupthink", a tendency to focus upon only a subset of the information that is available to all inspectors ("shared information bias") or it leads individuals to "free-ride" on the effort of others (Bang & Firth, 2017). Again, however, we know of little analogous evidence on the association between inspection team size and final inspection outcomes, including differences between group versus individual inspections. We thus conclude by asking:

RQ7. Do school inspection outcomes vary by inspection team size? Do outcomes differ between team versus individual inspections?

3. Data

Watchsted database

Our primary data source is drawn from the "Watchsted" website, which allows schools to look up details about Ofsted inspectors (https://perspective.angelsolutions.co.uk/Perspective/LiteUsers/Ofsted/InspectorSearch.aspx). For each inspector, this includes details of all inspections that they have conducted since September 2011[3], based upon the name of the lead inspector published within the written Ofsted reports (information which is in the public domain). We have extracted from the Watchsted database all secondary inspections and all primary inspections done by inspectors who have conducted at least five between September 2011 and August 2019 (our period of interest). These data have then been cleaned, leading to records for a small number of inspectors

[3] They have gathered this information based upon the name of the lead inspector published within the written Ofsted reports.

to be merged where a similar name is used (e.g. Ash Rahman and Ashfaq Rahman have been combined into a single record)[4].

The Watchsted database has then been merged with publicly available information published by Ofsted containing details on all inspection outcomes (https://www.gov.uk/government/statistical-data-sets/monthly-management-information-ofsteds-school-inspections-outcomes). This was done in three steps. First, for each inspection extracted from the Watchsted website, we take the reported inspection start date and restrict the data published by Ofsted to only those inspections that were conducted on that date (i.e. we force there to be an exact match on inspection start date). Second, within this subset, we fuzzy match across the two databases on school name. Finally, we check that the information on inspection outcomes – including sub-judgements – is consistent across the two sources. Cases were dropped in the few instances where differences were found. This process was conducted separately for primary and secondary schools before being combined.

The final dataset used in our analysis thus includes 35,751 inspections (29,850 primary and 5,901 secondary) conducted between September 2011 and August 2019 by a total of 1,376 inspectors. This represents 81% of all inspections conducted over this period, with Appendix B discussing this issue in further detail. Appendix B also provides alternative estimates based upon a different sample selection using data from 40,959 inspections (93% of the total). We find that this leads to little change in our substantive results.

Appendix C provides details about how we have checked the quality of the data we have extracted. In brief, we have randomly sampled 300 inspections, accessed the relevant inspection reports from the Ofsted website and manually recorded the relevant information (e.g. inspector name, whether an HMI led the inspection). We then cross-reference this information against what is recorded in our dataset to check their consistency. Overall, the level of agreement is high, with the name of the lead inspector the same on more than 95% of occasions. This, in turn, provides reassurance that measurement error in our data is likely to be low.

These data were then subsequently linked to further information about schools, drawn from the Department for Education's (DfE) School Performance Tables (https://www.compare-school-

[4] In such instances, we are making an assumption that this is the same individual, who is simply using a different variant of their name within the written inspection reports.

performance.service.gov.uk/download-data). This includes information about the background of schools (e.g. admissions policy, religious denomination, school type), the composition of the student body (e.g. percent of pupils eligible for Free School Meals, percent of pupils with English as an Additional Language) and performance in national examinations. Together, this provides key background information about the schools that inspectors inspected.

Based upon the information available within the database, we also derive the following about individual inspectors:

- Whether an HMI. For each inspector named in the Watchsted database, there is a flag to indicate whether they are an HMI. Any inspector with such a flag is coded as an HMI, with all others assumed to be an Ofsted Inspector (OI).
- Gender. The python GenderGuesser (https://pypi.org/project/gender-guesser/) package is used to predict the gender of each inspector, based upon their first name. A small amount of manual coding has also been conducted, where results from the GenderGuesser package were ambiguous.
- Primary/secondary specialism. Some inspectors conduct inspections in a single school phase (primary or secondary) while others work across both. We thus derive a variable, based upon each inspector's inspection history, identifying whether each inspector has conducted primary inspections only, secondary inspections only or done a mix of both.
- Home region. Ofsted operates a regional operating model, with each inspector sitting within one regional inspection team. It is possible however for inspectors to sometimes conduct inspections outside of their "home" region. For each inspector who has conducted more than 10 inspections between September 2011 and August 2019, we define their "home region" as the area where they have conducted most of their inspections[5]. Using this information, we also derive a binary variable, identifying for each inspection whether the inspector was working in their home region or not.
- Experience. Total inspection experience is measured as the number of inspections an inspector has previously conducted (before their current inspection) with the count starting in September 2011.
- Inspection team size. This is measured as the number of inspectors named as participating in the inspection. As this information is not available from the Watchsted

[5] Inspectors who have conducted more than half of their inspections outside of their "home" region have been recoded into a separate category of "no home region".

website we have extracted this information via our own scraping of the Ofsted reports (see Appendix B for further details).

A set of descriptive statistics, documenting the distribution of these variables across all inspections included within our analysis, can be found in Table 3. HMIs are slightly more likely than OIs to lead short inspections (60%/40% split). For other inspection types, however, OIs are more likely to be the lead than HMIs – particularly in primary schools (80%/20% split). This is an important point, given that – as noted in the introduction – most previous work into the reliability and consistency of Ofsted inspections has not included OIs. Despite women being more likely to work in the teaching profession than men - particularly in primary schools (Jerrim & Sims, 2019) - the same does not hold true with respect to inspections, where the gender split is broadly even. Most primary inspections are conducted by primary inspection specialists, although around 10% are led by an inspector whose workload has included a significant proportion of secondary inspections. The analogous holds true with respect to secondary inspections. While short inspections are almost always conducted within an inspector's home region, approximately one-in-seven (15%) of non-short inspections are conducted outside of it. The average primary inspection is led by someone who has led around 30 inspections previously, though there is quite a lot of variability around this figure (the standard deviation is approximately 25). For secondary inspections, the average amount of prior lead experience is somewhat lower (an average of 17 prior inspections led). Finally, primary inspections are conducted by smaller inspection teams. Almost two-thirds of primary inspections (that are not short inspections) are conducted by one or two inspectors (63%), compared to just 14% of secondary inspections. This will partly reflect differences in primary and secondary school size.

<< Table 3 >>

Outcome measures

Our primary outcome of interest is the Overall Effectiveness judgement awarded to schools (Outstanding, Good, Requires Improvement, Inadequate). However, we will also on occasion consider differences in outcomes on the various Ofsted sub-domains, including:

- Behaviour and safety of pupils (September 2011 – August 2015)
- Personal development, behaviour and welfare (September 2015 – August 2019)
- Leadership and management (September 2011 – August 2019)

- Outcomes for pupils (September 2015 – August 2019)
- Quality of teaching, learning and assessment (September 2015 – August 2019)
- Quality of teaching (September 2011 – August 2015)

Our outcome for short inspections is a binary measure, coded as one if the inspector decided the school should receive a full inspection next due to concerns, or for the conversion to a full inspection immediately with a subsequent downgrade in Overall Effectiveness judgement, and coded zero otherwise. Given the changes to short inspections in January 2018, we will at times also present results separately for the September 2015 – December 2017 and January 2018 – August 2019 periods.

4. Methodology

RQ1. What proportion of the variation in inspection outcomes occurs between (rather than within) inspectors?

To address research question 1 the following random effects (i.e. multi-level) models are estimated:

$$O_{ij} = \alpha + \mu_j + \varepsilon_i \quad \text{(Model 1a)}$$

$$O_{ij} = \alpha + \beta.X_i + \mu_j + \varepsilon_i \quad \text{(Model 1b)}$$

Where:

O_{ij} = Overall effectiveness inspection judgement.

X_i = A vector of inspection-specific controls (e.g. prior inspection rating, school type, inspection type, historic performance in national examinations).

μ_j = An inspector level random effect.

ε_i = An inspection-specific error term.

i = Inspection i.

j = Inspector j.

These models hence treat inspections (level 1) as nested within inspectors (level 2). The outcome (O_{ij}) is initially treated as ordinal, with the model estimated using ordinal logistic regression. This provides an overall indication of the extent that Ofsted inspection outcomes vary across different inspectors. However, we also estimate a set of models where O_{ij} has been

dichotomised (e.g. O_{ij} = 1 for an Inadequate rating and zero otherwise) to explore whether there are greater between-inspector differences for certain Ofsted grades.

The primary statistic of interest from these models is the intra-cluster correlation. This captures the proportion of the total variation in Ofsted outcomes that occurs between different inspectors. Formally, this is defined as:

$$\rho = \frac{\sigma_{\mu}^2}{(\sigma_{\mu}^2 + \sigma_{e}^2)}$$

Where:

ρ = The intra-cluster correlation.

σ_{μ}^2 = The variation in Ofsted judgements that occurs between different inspectors.

σ_{e}^2 = The variation in Ofsted judgements that occurs within inspectors.

Following the standard approach in the multi-level modelling literature, unconditional ICCs are reported from an "empty" model (a model that does not include any controls). Such estimates fail to recognise, however, that certain inspectors may be disproportionately assigned to certain inspections. We consequently also report conditional ICCs, where a range of inspection-specific covariates are controlled. Estimates are reported separately for primary/secondary phases. The analogous analysis is then replicated for our binary short-inspection outcome.

The link between inspection outcomes and observable lead inspector (or inspection team) characteristics (RQ 2-7)

A similar analytic process is followed for each of our observable lead inspector characteristics and for inspection team size.

To begin, we present simple unconditional descriptive statistics illustrating how each characteristic is related to inspection outcomes. Of course, these unconditional relationships may be confounded by other factors. In particular, Ofsted may assign inspectors with certain characteristics (e.g. those that they perceive to be of higher quality) to inspect certain types of school. For instance, Ofsted may disproportionately assign HMIs to schools where they think a difficult judgement might need to be made.

We consequently estimate a set of ordered logistic regression models to try and account for the possible differential selection of lead inspectors to different types of school. These models control for a set of factors known to be related to overall inspection outcomes and may be associated with inspector (and inspection team) assignment. All models will be estimated separately for primary and secondary schools, recognising the important differences between these different school phases. These models are of the form:

$$\log\left(\frac{P(O_{ij} \leq k)}{P(O_{ij} > k)}\right) = \alpha + \beta.I_j + \tau.X_i + \tau.C_J \qquad (1)$$

Where:

O_{ij} = Overall inspection judgements measured using Ofsted's four-point scale.

I_j = The characteristic of the lead inspector / inspection team under investigation.

X_i = A vector of inspection-specific controls. These are either characteristics of the school being inspected (e.g. performance in national examinations) or the type of inspection being conducted.

C_J = Other characteristics of the lead inspector (other than the characteristic under investigation).

i = Inspection i.

j = Inspector j.

k = A specific category on Ofsted's four-point overall effectiveness scale.

The parameter of interest from these models is β. This illustrates the strength of the association between the characteristic currently under investigation (e.g. gender) and overall inspection outcomes. Estimates will be presented in terms of odds ratios, capturing the increase in the odds of receiving a worse inspection rating. For instance, an odds ratio of two would indicate that the odds of receiving an Outstanding rating versus a Good/RI/Inadequate rating are twice as large, conditional upon the other factors controlled in the model. We will also on occasion present predicted outcomes (predictive margins) from our regression models to further aid interpretation of results.

Several different versions of this model will be estimated to investigate the robustness of our findings to the inclusion of different covariates. Our baseline specification (M0) will not include any controls. It will thus act as a benchmark against which estimates from the other model specifications can be judged. Model M1 will add a set of basic school background characteristics, such as the gender of pupils at the school, the school's religious denomination, Ofsted phase, percent of disadvantaged pupils, and region. We then add controls for inspection type (e.g. Requires Improvement re-inspection, exempt school inspection, S5 inspection) and prior inspection judgement. This model thus accounts for prior judgements Ofsted has made about the quality of the school (M2). Model M3 then adds controls for school's recent performance in national examinations (Key Stage 2 scores for primary schools and GCSEs for secondary schools). Further attributes of the school, including absences and percent of pupils with English as an Additional Language (EAL) or Special Educational Needs (SEN) are added in M4. Finally, models M5 and M6 add further characteristics of the inspector (C_J), over and above the specific characteristic under investigation (I_j). For instance, HMIs may be more likely to be male than female. Models M5 and M6 will take this into account, thus illustrating (for instance) whether male and female inspectors make different judgements about schools, taking into account that they may have different contracts (and thus working relationship) with Ofsted.

To account for the nested structure of the data, with inspections conducted within inspectors, standard errors will be clustered at the inspector (j) level. In Appendix D we also replicate some of our headline findings using multi-level (random effects) models to test the robustness of our results to an alternative analytic approach. On occasion, we will also estimate multinomial (rather than ordinal) logistic regression models, to investigate the sensitivity of our findings to relaxing the proportional odds assumption. Analogous models to those presented in equation (1) will be estimated for the sub-judgements (when considered) and for short inspection outcomes[6].

Joint effect – looking at the impact of multiple characteristics together

To investigate the combined effect of multiple inspector characteristics we estimate a multinomial logistic regression model[7] including our five lead inspector characteristics of

[6] As our outcome from short inspections is binary, these models are estimated using binary (rather than ordinal) logistic regression.
[7] Alternative estimates will also be presented using an ordinal logistic regression model.

interest at the same time (gender, HMI/OI, experience, inspecting outside of the home region and phase specialism) along with inspection team size and a set of school/inspection level controls[8]. From this model we predict the probability that two hypothetical inspectors (A and B) award each of the four Ofsted overall effectiveness judgements to a school. Specifically, we consider differences in the distribution of primary school inspection outcomes between the following two hypothetical inspectors:

- Inspector A. An inexperienced, female HMI who specialises in primary school inspections and who is undertaking an inspection outside of their home region. The inspection team size is set to two inspectors.
- Inspector B. An experienced, male OI who undertakes both primary and secondary inspections and who is currently completing an inspection within their home region. The inspection team size is set to one inspector.

This part of our analysis will focus upon differences in primary school inspection outcomes given the much larger sample size for this Ofsted phase. The results we report will be when these two hypothetical inspectors are inspecting schools with a similar proportion of disadvantaged pupils, within the same Ofsted region, have similar levels of performance in the Key Stage 2 tests, have the same previous Ofsted inspection judgement, have similar levels of school absence, similar proportions of pupils who speak English as an Additional Language and undergoing the same type of inspection.

5. Results

RQ1. What proportion of the variation in inspection outcomes occurs between (rather than within) inspectors?

Table 4 illustrates the percent of the variation in inspection outcomes that occurs between different inspectors. The top row ("ordinal") refers to where the four-point Ofsted overall effectiveness judgement is treated as an ordinal outcome. The next four rows present estimates from multi-level logistic regressions, where the outcome has been dichotomised for each overall effectiveness judgement in turn (e.g. the "Good" row refers to variance across lead inspectors in awarding Good compared to any other judgment).

[8] These are percent of pupils eligible for FSM, region, previous Ofsted inspection outcome, inspection type, Key Stage 2 maths and English scores, school absences, percent of pupils with English as an additional language and whether the inspection was conducted after 2018

<< Table 4 >>

For overall effectiveness grades of primary schools, around 10% of the variance in overall effectiveness judgements occurs across different lead inspectors (when treating this as an ordinal outcome). Interestingly, this result is largely unaffected by the addition of background school-level controls. The second to fifth rows indicate that this result is being mainly driven by differences across inspectors in awarding the top (Outstanding) and bottom (Inadequate) grades. In particular, more than 15% of the variation in awarding Inadequate grades occurs across different inspectors, suggesting that some inspectors are much more likely to award this high-stakes grade than others[9]. To put these figures into context, previous research has found a similar proportion of the variance in Key Stage 2 scores to occur between different primary schools (Allen et al., 2018: Table 3). In other words, the "clustering" of different inspection outcomes across different inspectors is similar to the variation in achievement outcomes amongst children who attend different primary schools.

The variation in Overall Effectiveness judgements across lead inspectors is somewhat lower for secondary schools, standing at around 7.4% in the unconditional model and 5.2% once background characteristics of the school have been controlled. The variation we observe across inspectors is clearly being driven by differences in their propensity to reach the highest (Outstanding) and lowest (Inadequate) inspection judgements. For instance, from the conditional models, one can see that more than 10% of the variance in Inadequate and Outstanding grades occurs between different inspectors, compared to little more than 2% for the Good and Inadequate grades.

Finally, the bottom row presents analogous estimates for short inspection outcomes. In the unconditional models, around 12% of the variation in primary short inspection outcomes occurs between inspectors, falling to around 11% once school background characteristics have been controlled. This is broadly consistent with our analysis of overall effectiveness judgements outlined above, in that there appears to be non-trivial differences across inspectors, even after background characteristics of the schools they have inspected have been controlled. On the other hand, the between-inspector variation is notably smaller for short inspections of secondary schools, standard at 5% in the unconditional model and 0% in the conditional model.

[9] We note, however, that many inspectors may never award an inadequate grade at any point during their inspection career; particularly those who have conducted comparatively few inspections.

RQ2. Do female inspectors make harsher lenient judgements about schools than their male counterparts?

Table 5 begins by presenting the distribution of overall effectiveness judgements made by male and female lead inspectors.

<< Table 5 >>

Starting with primary schools, evidence emerges of a modest (though not trivial) gender difference. Female lead inspectors seem to reach somewhat harsher judgements about primary schools than their male counterparts. For instance, male lead inspectors judged 33.1% of primary schools to Requires Improvement or to be Inadequate, compared to 36.4% of female lead inspectors. The difference in the Inadequate grade (5.9% versus 4.5%) is particularly notable, given the size of the relative gender difference, and the high-stakes consequences of such a decision being made. Male lead inspectors are, on the other hand, more likely to judge schools to be Good than their female counterparts, where a 2.9 percentage point difference emerges. There is little evidence of a gender gap, however, in the awarding of an Outstanding grade. Nevertheless, Table 5 provides a first suggestion that the inspection outcome of primary schools may to some extent be influenced by the gender of the lead inspector.

The results for secondary schools – presented on the right-hand side of Table 5 – are not as clear. The percentage of male and female lead inspectors awarding Good and Requires Improvement grades are very similar. There is perhaps more of a difference at the extremes of the grading scale, with male lead inspectors more likely to reach an Inadequate judgement (10.5% versus 9.1%) and female leads more likely to award Outstanding grades (10.9% versus 10.1%). Yet even these differences are relatively small, and hence the evidence being mixed, at best.

To what extent might these unconditional results be driven by "selection"; are the apparently hasher judgements made by female inspectors due to them being assigned more challenging primary schools? Two pieces of evidence are presented on this matter. First, Table 6 compares the distribution of observable school-inspection characteristics between male and female lead inspectors. This is important as, if female lead inspectors are indeed assigned to inspect lower-quality schools, one would expect to see female inspectors being disproportionately assigned to schools with lower prior inspection ratings, worse performance in national examinations or

higher absence levels. Yet there is no evidence that this is that case; the distribution of inspection tasks appears very similar across male and female lead inspectors.

<< Table 6 >>

Second, Table 7 presents estimates from a set of ordinal regression models[10]. Odds-ratios are reported, with values below one indicating that female lead inspectors make harsher judgements than their male counterparts.

<< Table 7 >>

Model specification M0 presents the unconditional estimates, and thus reflects the descriptive pattern observed previously. The estimated odds-ratio is 0.86, and is statistically significant at the five percent level, reiterating that female lead inspectors tend to award lower inspection grades to primary schools than male inspectors. Models M1-M4 then adds a series of inspection-level controls capturing different aspects of the inspected schools. This includes their demographic composition, performance in national examinations (i.e. school-level average Key Stage 1 and 2 scores), inspection type and how they were judged during their previous inspection. If inspector selection were driving the pattern observed in Table 5, one would expect the estimated odds-ratio to get closer to one between models M0 and M4. This is clearly not the case. Indeed, the odds ratio slightly *decreases* between model M0 (0.86) and M4 (0.82) suggesting that – if anything – accounting for inspector selection may strengthen our result. Moreover, the parameter estimates for lead gender inspector also do not substantively change in models M5 and M6 when other observable characteristics of inspectors are added into the model.

Although we can of course only control for observable characteristics, the stability of the estimated odds-ratio across model specifications indicates that any unobserved confounding would have to be generated by a factor that is strongly associated with inspection outcomes, but also be orthogonal to a school's intake, performance in examinations, pupil absences and previous Ofsted grades. It is not clear what such a characteristic could be. Our interpretation is hence that results presented above provides strong evidence that the gender difference we observe in primary inspection outcomes is not being driven by inspector selection.

[10] Analogous results for secondary schools are provided in Appendix G. These confirm the finding that for secondary schools there is no clear link between the gender of the lead inspector and inspection outcomes.

Do female lead inspectors make harsher judgements across the board? Or are there particular aspects of primary schools that they rate lower than their male peers? We consider this issue in Table 8 where we replicate our analysis (using model specification M6) for each of the Ofsted sub-judgements. This provides little evidence that any area stands out. The estimated odds-ratios are very similar for each of the sub-scales, typically falling between 0.80 and 0.85. It hence seems that female inspectors generally rate primary schools to be lower quality than male inspectors, with this not seemingly being driven by differences in opinion on one specific area.

<< Table 8 >>

Might our results be driven by a small number of inspectors or by a particular sub-group? We explore this possibility in Appendix E by re-estimating our ordinal logistic regression model for different sub-groups. Our results remain unchanged if we focus upon just HMIs or just OIs, or if we restrict the sample to include S5 inspections only. There is also no clear pattern of the results varying by academic year, though the vastly reduced sample sizes means these estimates are somewhat noisy and fluctuate in terms of their statistical significance. A similar finding holds true with respect to geographic region. This leads us to conclude that there is no evidence that the gender difference we observe in primary inspection outcomes is driven by a specific sub-group of inspectors.

We have also re-estimated our analytic models using multinominal (rather than ordinal) logistic regression. These estimates can be found in Appendix F. Our substantive conclusions once again remain unchanged. In particular, they confirm that there is little evidence of a gender difference when it comes to the Good/Outstanding distinction (consistent with the descriptive cross-tabulation presented in Table 5). Rather, the difference seems to be driven by differences in male and female lead inspectors reaching different Good/RI/Inadequate judgements.

To conclude, Table 9 turns to the association between lead inspector gender and short inspection outcomes. In particular, it focuses upon the chances of a negative outcome from such an inspection defined as:

(a) Pre January 2018. Conversion to a full inspection with an Overall Effectiveness judgement of Requires Improvement or Inadequate being made.
(b) Post January 2018. Immediate conversion to a full inspection with a subsequent downgrade or recommending an S5 inspection be conducted next due to concerns.

The estimates presented in Table 9 are based upon a logistic regression model that includes a wide array of background school and inspector controls[11]. Odds ratios below one indicates that short inspections with a male lead are less likely to result in a negative outcome than their male counterparts.

<< Table 9 >>

Panel (a) presents results for short primary inspections. This suggests that male leads tend to reach less harsh judgements from short inspections than female leads. The estimates odds ratios sit around 0.8, indicating that the odds of a poor outcome from a short inspection are around 20% lower for males than females. In absolute terms, this represents a modest difference of around two percentage points; the chances of a negative outcome with a male lead inspector are around 11.5%, compared to 13.5% for a female lead. There is, on the other hand, no evidence of such a gender difference with respect to short secondary inspections. Hence, overall, our findings with respect to short inspection outcomes are consistent with those from full inspections. In particular, both point towards a small lead inspector gender difference in inspection outcomes – at least for primary schools.

RQ3. Do Ofsted inspection judgements differ between OIs and HMIs?

Table 10 illustrates the distribution of inspection outcomes by inspector contract status (HMI versus OI). Starting with the results for primary schools, one can see a clear difference between the two groups. Most notably, HMIs are around 13 percentage points less likely to award a Good grade than OIs (60% versus 47%) but are much more likely to judge schools to require improvement (36% versus 28%) or to be Inadequate (9% versus 4%). Contract status does thus seem to be related of Ofsted judgements reached, at least descriptively.

<< Table 10 >>

Evidence of a difference between OI and HMIs for secondary schools is somewhat more mixed. The percentage of secondary schools awarded an Outstanding or Requires Improvement grade is very similar across the two groups. Where evidence of a difference emerges is with respect to the Good and Inadequate judgements. Specifically, HMIs judge fewer schools to be Good than OIs (43% versus 47%) but place more in the Inadequate category (12% versus 8%).

[11] Appendix H presents alternative estimates based upon different model specifications and sample selections. Although the statistical significance of estimates fluctuates between the 10% and 5% level, the estimated odds ratio remains broadly stable.

Nevertheless, the magnitude of the difference between HMI and OI lead inspectors seems to be greater at primary than secondary level.

Next, we investigate the relationship between inspector contract status and inspection outcomes more formally, via estimation of a series of ordinal logistic regression models. The estimates for primary schools can be found in Table 11, with odds-ratios greater than one indicating that HMIs tend to provide harsher inspection judgements than OIs.

<< Table 11 >>

There are two key points to note. First, the relationship between contract status and inspection outcomes is strong and statistically significant across all model specifications. Roughly speaking, the odds of a primary school being placed in a lower Overall Effectiveness category is around 50% higher if the lead inspector is an HMI rather than an OI. Second, the inclusion of various inspection, school and inspector controls only leads to a slight weakening of the relationship. Across the model specifications, the estimated odds ratio consistently sits between around 1.4 and 1.5. This, in turn, indicates that this result is not being driven by the selection of HMIs/OIs into different types of inspection, at least in terms of a set of key observable characteristics (such as examination performance and demographic composition). We cannot rule out the possibility, however, that HMIs and OIs are disproportionately chosen to conduct inspections based upon a factor we cannot observe (and is not well-proxied by our wide range of controls).

Analogous results for secondary schools are presented in Table 12. These confirm that the relationship between inspector contract status is weaker amongst secondary schools than their primary counterparts; being inspected by an HMI is associated with only around a 20% increase in the odds of being awarded a lower inspection grade. The addition of control variables further weakens the relationship – most notably the inclusion of inspection type and prior Ofsted rating. However, alternative estimates based upon multinominal (rather than ordinal) logistic regression in Appendix F make clear that for secondaries, the main point of difference between HMIs and OIs is with respect to the Good and Inadequate grades. This difference, it seems, is largely unaffected by the addition of various controls.

<< Table 12 >>

Table 13 turns to estimates for the separate Ofsted sub-judgements. For primary schools we observe a difference in the judgements made by both OIs and HMIs in each of the sub-domains.

The estimated odds ratios are particularly large for behaviour (odds ratio = 1.64) and teaching (odds ratio = 1.60), with those for the other judgements all around 1.4 and below. However, on the whole, HMIs seem to judge primary schools more harshly in most areas.

For secondary schools, estimates for most of the domains are around one and fail to reach statistical significance at conventional levels. There are, however, two notable exceptions. The most prominent is pupil behaviour, where the odds of an HMI awarding a lower grade being around 50% higher than for OIs. A similar pattern holds for teaching, where the odds are 30% higher. Together this suggests that, at secondary level, HMIs and OIs may differ in their views of specific aspects about the quality of a school.

<< Table 13 >>

Appendix E presents separate results by sub-group. At primary level, there is no clear evidence that the aforementioned findings have notably changed over time. There is also no change to our results when we restrict the analysis to Section s inspections only. In terms of regional variation, the odds ratio for the South West (2.23) stands out as notably larger than for the other areas (i.e. in the South West, differences in inspection outcomes between HMIs and OIs are particularly pronounced).

On the other hand, for secondary schools there has been a marked change over time. Any difference between OIs and HMIs seems to be driven – for secondaries – by results before September 2015 (the start of the 2015/16 academic year). This coincides with significant changes at Ofsted, including (a) the introduction of short-inspections (see below for further analysis) and (b) Ofsted inspectors becoming directly contracted by Ofsted, rather than being outsources to a third party. The latter also meant that the pool of Ofsted inspectors decreased substantially (from 3,000 to around 1,600) through a selection process (https://www.bbc.co.uk/news/education-33198707). Otherwise, the only other notable point for secondary schools is that the only region where there is a statistically significant difference between HMIs and OIs is London (due partly, however, to the limited regional level sample size).

To conclude, we consider HMI and OI differences in short inspection outcomes. Table 14 presents the percent of short inspections conducted by OIs and HMIs in the primary and secondary sectors. This reveals how the short inspection is more likely to lead to a bad outcome (conversion with a downgrade or recommendation of a full S5 inspection next due to concerns)

if its conducted by an HMI rather than an OI (14% versus 10%). A similar difference holds for secondary schools (18% versus 22%).

<< Table 14 >>

Table 15 adds further detail to this analysis by presenting logistic regression model estimates, capturing the difference between HMIs and OIs in a short inspection leading to a negative outcome (conditional upon other background factors of the school being controlled). This reveals that HMIs are more likely to make a negative judgement than OIs when conducting a short inspection. The estimated odds ratio for both primary and secondary inspections is around 1.4, suggesting that the odds of a negative outcome is around 40% higher for HMI led inspections (compared to OI led inspections). Interestingly, the estimated odds ratio for short primary inspections is higher after the January 2018 changes were made than before (1.63 versus 1.27), although this difference is not statistically significant at conventional levels.

<< Table 15 >>

RQ4. How are inspection outcomes linked to inspection experience of lead inspectors?

Table 16 presents results from ordinal regression models estimating the relationship between inspector experience and overall effectiveness judgements. Results for both primary and secondary schools tell a similar story – there is no clear relationship between inspector experience and overall inspection grades. Most estimates sit close to one and are not statistically significant at conventional levels. This holds true regardless of the school, inspection or inspector controls included in the model. Similar results also emerge for short inspections, with little evidence of a clear, consistent link between inspector experience and a negative outcome.

<< Table 16 >>

RQ5. Do inspectors judge schools more harshly when they are working outside of their home region?

Table 17 begins by presenting cross-tabulations between whether the inspection was conducted inside the inspector's home region and the inspection outcome. For primary schools (panel a) the distribution of overall effectiveness judgements is very similar whether the inspection was conducted within inspector's home region or not. For secondary schools, however, it sems that inspections conducted outside of the home region leads to slightly better inspection grades.

Specifically, secondary inspections outside of the home region are more likely to be rated Outstanding (15 versus 9 percent) and slightly less likely to receive an Inadequate grade (7 versus 11 percent). Although results are also presented for short inspections, the number conducted outside of an inspector's home region is small, and thus not commented upon further here.

<< Table 17 >>

In Tables 18 and 19 we investigate whether these unconditional results continue to hold when we control for other factors within a set of ordinal regression models. For primary schools, the odds ratio in the unconditional model (M0) sits close to one (0.99). The addition of controls – particularly whether the inspector is an HMI, inspection type and prior inspection rating – drives the odds ratio upwards. In the final specification the estimated odds ratio (1.13) is statistically significantly above one, though modest in terms of magnitude. Overall, we thus conclude that evidence of a relationship between whether an inspection is conducted within the inspector's home region and overall effectiveness judgements is weak.

<< Table 18 >>

With respect to secondary schools the opposite holds true. In the unconditional model (M0) the estimated odds ratio is significantly below one (0.73). This apparent relationship quickly disappears however in M1 once basic background controls about the school being inspected are added to the model (odds ratio increases to 0.92). The addition of further controls does little to change this result, with the estimated odds ratio in the most detailed model specifications sitting almost exactly on one. We thus conclude that there is no evidence that the inspection judgement secondary schools receive is related to whether the lead inspector was working in their home region or not.

RQ6. Do inspectors with a specialism in secondary school inspections judge primary schools more harshly than inspectors with a primary specialism?

Table 20 panel (a) presents a cross-tabulation between the percent of primary school inspections each inspector conducted throughout their career and inspection outcomes. With respect to Overall Effectiveness judgements, there is no clear relationship. This largely continues to hold even after controlling for a set of school, inspection and inspector characteristics within our set of ordinal regression models (see Table 21), with only a slight difference emerging between the 30-69% and 100% primary groups (odds ratio = 0.86 –

statistically significant at the 10% level). The results are similar for secondary schools. The initial cross-tabulation (see Table 20 panel b) does not show evidence of a clear pattern between inspector phase specialism and overall effectiveness outcomes, which is then supported by results from our ordinal logistic regression models (see Table 22). All-in-all, it therefore seems that there is little evidence of an association between whether inspectors specialise in primary/secondary inspections and the overall effectiveness judgements reached.

<< Table 20 >>>

<< Table 21 >>>

<< Table 22 >>>

The same does not appear to be true, however, with respect to outcomes from short inspections (the probability of converting a short inspection or recommending an S5 inspection next). In the initial cross-tabulation, there is some suggestion that primary/secondary "specialists" (i.e. inspectors who only ever conduct inspections within one phase) are less likely to convert an inspection than non-specialists. This pattern is clearest within the secondary sector, where the chances of a negative outcome from the short inspection is 18% for secondary specialists versus 25% for non-specialists.

At least for secondary schools, these patterns largely remain intact throughout our logistic regression modelling process (see Tables 23 and 24). Indeed, for secondary schools, the addition of controls for school, inspection and other inspector characteristics seems to further strengthen the association, from odds ratios around 1.5 in M0 (unconditional model) to around 2.0 in model M8 (full specification). Findings from these models are summarised in Table 25, where we use the estimates to predict the probability of a short inspection conversion / S5 recommended next by phase specialism. Consistent with results from the initial cross-tabulation, it seems that specialists are less likely to convert short inspections than non-specialists – with the clearest evidence coming in the secondary sector. For instance, amongst secondary schools with a similar demographic intake, similar performance in national examinations and similar levels of school absences, an inspector with a secondary specialism (i.e. only ever inspected secondary schools during their inspection career) has a 19% chance of it resulting in a bad inspection outcome for the school, compared to more than a 26% chance for a non-secondary specialist.

<< Table 23 >>>

<< Table 24 >>>

<< Table 25 >>>

RQ7. Do school inspection outcomes vary by inspection team size? Do outcomes differ between team versus individual inspections?

Table 26 presents summary statistics for the link between inspection team size and inspection outcomes. For primary inspections, there is a clear trend between team size and inspection outcomes; larger teams are less likely to reach a Good judgement and are more likely to rate schools as Inadequate or Requires Improvement. The analogous association for secondary inspections is somewhat less clear, particularly given the smaller sample available. With respect to short inspections, there seems to be a clear association between team size and unfavourable outcomes pre January 2018. However, post 2018 this link is less clear (particularly for primary schools). One possible explanation is that when a short inspection immediately converts to a full inspection, all inspectors are named in the report (those involved in the initial short inspection as well as those involved in the full inspection). As conversion from short to full inspections was a lot more common before January 2018 (see Table 2) this is likely to explain the substantial difference in results for short inspections conducted pre/post 2018.

<< Table 26 >>>

Tables 27 investigates whether the descriptive pattern observed for primary inspections continues to be observed once various school and inspector characteristics are controlled in a set of ordinal logistic regression models. Starting with the results for primary schools, the estimated odds-ratios fluctuate slightly across the model specifications. However, the difference between a single inspector versus inspections teams with two or three inspectors is consistently statistically significant, with the estimated odds-ratios typically between 1.2 and 1.3. Indeed, the estimated odds ratios for team sizes of two and three in model M6 (full set of controls) are little different from those in model M0 (no controls). Additional multinominal logistic regression estimates again point towards the most notable difference to occur with respect to the Inadequate grade. Specifically, the predicted probability of receiving an Inadequate grade is 3.4% when the primary inspection is conducted by a single inspector,

versus around 6% when it is conducted by a team of two, three or four inspectors[12]. See Appendix F for further details.

<< Table 27 >>>

Table 28 provides analogous estimates for secondary schools (although note that the reference category is now set to a team size of four inspectors which, as illustrated by Table 25, is the modal category for secondary schools). Interestingly, very small teams (one inspector) and large teams (five inspectors or more) seem to make less slightly less harsh judgements than secondary inspections conducted by a team of four. The estimates for these two categories are consistently statistically significant at the five percent level, with the estimated odds ratio around 0.5 with respect to a single inspector (relative to a team of four inspectors) and 0.8 for a team of five inspectors. There is hence some evidence that – for both primary and secondary inspections – inspection team size remains independently associated with Ofsted inspection outcomes, over and above our school and inspection level controls.

<< Table 28 >>>

Joint effect – looking at the impact of multiple characteristics together

To conclude, we examine the combined effect of multiple inspector characteristics. To do so, we estimate a multinomial logistic regression model including the three characteristics we have found to be most clearly associated with inspection outcomes of interest at the same time (gender, HMI/OI and team size) along with a rich set of school/inspection level controls[13]. From this model we predict the probability that two hypothetical inspectors (A and B) award each of the four overall effectiveness judgement. Specifically, we consider differences in the distribution of primary school inspection outcomes between the following two hypothetical lead inspectors:

- Inspector A. A female HMI working with one other inspector (team size =2).
- Inspector B. A male OI working alone (team size = 1).

[12] Based upon a model controlling for percent of pupils eligible for FSM, previous inspection rating, inspection type, school performance measures, school absences and whether the lead inspector is an HMI.
[13] These are percent of pupils eligible for FSM, region, previous Ofsted inspection outcome, inspection type, Key Stage 2 maths and English scores, school absences, percent of pupils with English as an additional language and whether the inspection was conducted after 2018

We focus upon differences in primary school inspection outcomes given the much larger sample size and (consequently) the stronger evidence that has emerged of differences in inspection outcomes in the preceding sub-sections. Note that the results we report are for when these two hypothetical inspectors are inspecting schools with a similar proportion of disadvantaged pupils, within the same Ofsted region, have similar levels of performance in the Key Stage 2 tests, have the same previous Ofsted inspection judgement, have similar levels of school absence, similar proportions of pupils who speak English as an Additional Language and are conducting the same type of inspection.

Results can be found in Table 29. There is a clear, sizeable difference in the inspection outcomes reached by our two hypothetical lead inspectors. In particular, note how inspector A is estimated to be around four times more likely to award an Inadequate judgement than inspector B (13.4% versus 3.4%). Likewise, around half of the primary schools inspected by inspector A will be judged to be Inadequate or Requires Improvement, compared to around a third of primary schools inspected by inspector B.

<< Table 29 >>

The middle panel of Table 29 provides analogous estimates for the probability of a negative outcome from a short primary inspection. These focus on those conducted during the January 2018 – August 2019 period[14]. This again suggests that inspector A reaches harsher judgements than inspector B, with the probability of converting to a full inspection being 16% compared to 10%. In other words, inspector A is 1.6 times more likely to ask for the follow-up full inspection of the school than inspector B.

6. Conclusions

School inspections are now a common feature of education systems across the globe. Ofsted – the school inspectorate in England – is one example where a team of inspectors make high-stakes judgements about schools, which can lead to job losses amongst senior staff, influences parental selection of schools and can even lead to school closures (Eyles & Machin, 2019; Bokhove et al., In Press). Yet little previous research has been conducted into the reliability, consistency and variation in Ofsted inspection outcomes. The handful of studies that do exist tend to be limited in scope and scale and have been conducted by school inspectorates

[14] For the September 2015 – December 2017 period, inspection team size is captured post any conversion from short inspections made. We hence focus on short inspections conducted between January 2018 and August 2019.

themselves (e.g. Matthews et al., 1998; Ofsted, 2017a). Consequently, even some basic facts are currently unknown, such as how inspection outcomes vary across different inspectors.

This paper has sought to address this gap in the literature. Drawing upon data from more than 30,000 school inspections conducted over an eight-year period, we have produced – to our knowledge – the first published evidence on how school inspection outcomes are linked to characteristics of the lead inspector. Our findings suggest that almost 10% of the variation in primary school inspection judgements occur across different inspectors; this is similar to differences in achievement that occur across primary schools. Robust evidence emerges that male inspectors make more lenient judgements about primary schools than female lead inspectors, with this finding appearing unlikely to be due to "selection" (i.e. male and female lead inspectors being assigned to different tasks). Although the magnitude of such gender differences are relatively small, they are particularly pronounced at the highest stakes (Inadequate) grade.

We can only speculate as to why we observe the small but important gender differences in inspection judgements. One possibility is that the gender gap is being driven by differences in personality traits, with men being more likely to be overconfident in their knowledge and skills (Jerrim, Parker, & Shure, 2019), while women have higher levels of conscientiousness (Verbree et al., 2022). In job promotions, Hartman et al. (1991) argued that it's "predominantly the gender stereotype of the ratee's personal characteristics rather than the ratee's gender that influences the promotion process" (p. 285). In any case, it is plausible that such personality traits are linked to school inspection outcomes, thus driving the gender difference that we observe. Alternatively, previous research has suggested that there are important gender differences in decision making processes when working as part of a team. For instance, Kennedy (2003) notes how women tend to be more altruistic in their decision-making and prefer reaching a universal solution, while men are more motivated by self-interest. This could lead men and women to make different (high stakes) decisions, such as the inspection judgement awarded to a school. Villanueva-Moya and Exposito (2021) highlight the relevance of psychosocial variables like stereotype threat and fear of negative evaluation, in women's decision-making processes. Some evidence points towards effective interventions for stereotype threat (Liu et al., 2021), although some scholars argue that this depends on the form of stereotype threat (e.g. Shapiro et al., 2013). Finally, male and female inspectors may differ in their professional experiences, including their subject/phase specialisms and the leadership roles that they have held. Again, such factors may also be related to inspection outcomes, and

thus are also potential explanations for the gender difference we observe. Ultimately, however, this is an empirical question – and one that do we not have the data to answer. An important direction for future research is hence to develop a better understanding of what exactly is driving the gender difference in primary school inspection outcomes.

Much larger differences are observed between inspectors working under different contractual arrangements (HMIs versus OIs), with the former consistently reaching harsher judgements than the latter, even after controlling for a wide array of school and inspection characteristics. Likewise, inspection team size also appears to be independently associated with inspection outcomes, most notably with inspections being conducted by a single individual being less likely to lead to a negative outcome (and more likely to award the modal Good grade) than a team of two inspectors or more. On the other hand, little association was found between inspection outcomes and the lead inspector's experience, primary/secondary specialism or whether the inspection was conducted outside their home region. Likewise, partly due to the smaller sample size (and potentially also the bigger average inspection team size), weaker and more uncertain evidence of variation by lead inspector characteristics has emerged for secondary schools (in comparison to primary schools).

These findings should of course be interpreted considering the limitations of our work. Three issues stand out. First, our estimates capture conditional associations only, rather than capturing cause and effect. Some of the differences in outcomes we observe (e.g. between HMIs and OIs) may to some extent be driven by selection (different lead inspectors being assigned to different tasks). We have discussed this issue at length throughout the paper and have attempted to control for such differences in inspector deployment via estimation of various regression models. Nevertheless, we recognise that this may have only partially overcome such issues. Second, we have only considered variation by a limited set of key observable characteristics. Yet, arguably, the more important factor(s) driving the variation in inspection outcomes across inspectors is due to things we cannot observe within the data currently available, such as personality characteristics and traits. Further exploration of such wider characteristics should be a key line of enquiry for future research. Finally, a new inspection framework was introduced by Ofsted in September 2019, which puts less emphasis on data capturing performance in national examinations and more on the quality of the curriculum. Unfortunately, only six months of inspection data are available from this new framework before the COVID-19 pandemic hit England, with school inspections facing various forms of disruption over the following two years. Our analysis has thus been restricted to before the

most recent framework change. However, given that our analytic sample covers an eight-year period during which multiple changes were made to how school inspections were conducted (including changes to the framework) we have little reason to suspect that different findings would emerge now. Indeed, variation in judgements across inspectors may be greater now than under previous frameworks, given the move towards inspectors making professional judgements about curriculum quality, with less emphasis put upon more objective national examination data. Nevertheless, once data from further inspections are available under the new framework (outside of the pandemic era) we believe it important that Ofsted publish an update building upon our work.

With these caveats in mind, the key question becomes how much should our results be cause for concern? After all, Ofsted inspection frameworks explicitly recognises that inspectors should use their professional judgements when interpreting the evidence collected, with the variation we observe in our results merely reflecting this. In other words, there will of course be some degree of variation in outcomes in any process that involves human judgement. The most pertinent question thus becomes how much variation in outcomes across different inspectors is too much? This is obviously not a simple question to answer, and is itself open to discussion, opinion and debate. That said, we note that one of the clearest points of difference across lead inspectors in our work is with respect to what is widely perceived to be the highest stakes Inadequate grade. Given the consequences of receiving an Inadequate judgement, almost any variation across inspectors in reaching this decision might well be considered a problem.

What then should be the next step for Ofsted and other school inspectorates? Given the dearth of evidence on this matter – across the UK and internationally – school inspectorates should publish more research into the reliability and consistency of inspections, including variation in inspection outcomes. Only with such evidence at hand can an open and informed debate be had about such issues. At the same time, open data sources should also be created by school inspectorates – such as depositing in the Office for National Statistics Secure Research Service an inspector-inspection linked database – to allow independent researchers to also explore such issues. Likewise, more needs to be documented, investigated and discussed about inspector deployment – how exactly are inspectors assigned to different tasks? Finally, Ofsted might consider reviewing (and publishing) how it quality assures awards of an Inadequate grade (most notably those awarded to primary schools). Minimizing variation across inspectors in the chances of such an important, high-stakes judgement being reached – and explaining to the education sector how this is done – would be widely welcomed.

References

Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). Properties of commercial tests in the EEF database. Accessed 08/11/2021 from https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/methodological-research-and-innovations/Research_Paper_1_-_Properties_of_commercial_tests.pdf

Bang, D., & Frith, C. (2017). Making better decisions in groups. *Royal Society Open Science,* 4, 170193. https://doi.org/10.1098/rsos.170193

Bokhove, C., Jerrim, J., & Sims, S, (in press) How useful are Ofsted inspection judgements for informing secondary school choice? *Journal of School Choice.*

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In: *Teaching for Excellence and Equity. IEA Research for Education (A Series of In-depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA)), vol 6*. Springer. https://doi.org/10.1007/978-3-030-16151-4_2

Charness, G., & Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, *26*(3), 157–76. https://doi.org/10.1257/jep.26.3.157

Eyles, A., & Machin, S. (2019). The introduction of academy schools to England's education. *Journal of the European Economic Association*, *17*(4), 1107–1146. https://doi.org/10.1093/jeea/jvy021

Gernreich, C., & Exner, C. (2015). A comparison of the influence of gender on managerial decision making. Accessed 10/01/2022 from https://www.researchgate.net/publication/278679030_A_Comparison_of_the_Influence_of_Gender_on_Managerial_Decision_Making

Grohnert, T., Meuwissen, R., & Gijselaers, W. (2019). Enabling young professionals to learn from errors – the role of a supportive learning climate in crossing help network boundaries. *Vocations and Learning, 12*, 217–243. https://doi.org/10.1007/s12186-018-9206-2

Grund, C, & Thommes, K. (2017). The role of contract types for employees' public service motivation. *Schmalenbach Business Review, 18*, 377–398. https://doi.org/10.1007/s41464-017-0033-z

Hartman, S. J., Griffeth, R. W., Crino, M. D., & Harris, O. J. (1991). Gender-based influences: The promotion recommendation. *Sex Roles*, *25*(5), 285–300. https://doi.org/10.1007/BF00289757

Ingersoll, R. (1993). Loosely coupled organizations revisited. *Research in the Sociology of Organizations, 11*, 81–112.

Jerrim, J., & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766–781. https://doi.org/10.1093/esr/jcu072

Jerrim, J., Parker, P., & Shure, N. (2019). Bullshitters. Who are they and what do we know about their lives? IZA discussion paper 12282. Accessed 03/01/2023 from https://www.iza.org/publications/dp/12282/bullshitters-who-are-they-and-what-do-we-know-about-their-lives

Jerrim, J., & Sims, S. (2019). The Teaching and Learning International Survey (TALIS) 2018 report for England. Department for Education Research Report. Accessed 18/02/2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/919064/TALIS_2018_research.pdf

Kemethofer, D., Gustafsson, J-E., & Altrichter, H. (2017). Comparing effects of school inspections in Sweden and Austria. *Educational Assessment, Evaluation and Accountability,* 29, 319–337. https://doi.org/10.1007/s11092-017-9265-1

Kennedy, C. (2003). Gender differences in committee decision-making. *Women & Politics,* 25(3), 27–45. https://doi.org/10.1300/J014v25n03_02

Liu, S., Liu, P., Wang, M., & Zhang, B. (2021). Effectiveness of stereotype threat interventions: A meta-analytic review. *Journal of Applied Psychology*, *106*(6), 921–949. https://doi.org/10.1037/apl0000770

Matthews, P., Holmes, J. R., Vickers, P, & Corporaal, B. (1998). Aspects of the reliability and validity of school inspection judgements of teaching quality. *Educational Research and Evaluation, 4*(2), 167–188. https://doi.org/10.1076/edre.4.2.167.6959

McManus, I., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education, 6*, 42. https://doi.org/10.1186/1472-6920-6-42

National Education Union (2021). Abolish Ofsted and League Tables. Accessed 17/02/2022 from https://neu.org.uk/press-releases/abolish-ofsted-and-league-tables

Ofsted (2012). The framework for school inspection. Guidance and grade descriptors for inspecting schools in England under section 5 of the Education Act 2005, from January 2012. Accessed 17/02/2022 from https://dera.ioe.ac.uk/14077/1/The_framework_for_school_inspection_from_January_2012%5B1%5D.pdf

Ofsted (2017a). Do two inspectors inspecting the same school make consistent decisions? A study of the reliability of Ofsted's new short inspections. Ofsted research report 170004. Accessed 28/10/2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/596708/Reliability_study_-_final.pdf.

Ofsted (2017b). Ofsted confirms new arrangements for short inspections. Accessed 17/02/2022 from https://www.gov.uk/government/news/ofsted-confirms-new-arrangements-for-short-inspections

Ofsted (2017c). Ofsted seeking views on improved approach to short inspections. Accessed 18/02/2022 from https://www.gov.uk/government/news/ofsted-seeking-views-on-improved-approach-to-short-inspections

Ofsted (2019a). Workbook scrutiny. Ensuring validity and reliability in inspections. Ofsted research report 190028. Accessed 28/10/2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936240/Inspecting_education_quality_workbook_scrutiny_report.pdf.

Ofsted (2019b). How valid and reliable is the use of lesson observation in supporting judgements on the quality of education. Ofsted research report 190029. Accessed 28/10/2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936246/Inspecting_education_quality_Lesson_observation_report.pdf.

Ofsted (2019c). *Retaining the current grading system in education: some arguments and evidence*. Ofsted research report 190012. Accessed 24/08/2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936220/Retaining_the_current_grading_system_-_arguments_and_evidence_290419.pdf

Ofsted (2019d). Education inspection framework. Overview of research. Accessed 17/02/2022 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachme

nt_data/file/963625/Research_for_EIF_framework_updated_references_22_Feb_2021.pdf
Ofsted (2020). Ofsted inspections illustrate high proportion of Good or Outstanding schools. Accessed 11/01/2022 from https://educationhub.blog.gov.uk/2020/10/30/ofsted-inspections-illustrate-high-proportion-of-Good-or-Outstanding-schools/
Ofsted (2022a). School inspection handbook. Accessed 22/06/2022 from https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook
Ofsted (2022b). Inspecting schools: guide for maintained and academy schools. Accessed 17/02/2022 from https://www.gov.uk/guidance/inspecting-schools-guide-for-maintained-and-academy-schools#:~:text=The%20inspection%20will%20normally%20last,normally%20last%20for%201%20day.
Paloniemi, S. (2006). Experience, competence and workplace learning. *Journal of Workplace Learning*, *18*(7/8), 439–450. https://doi.org/10.1108/13665620610693006
Pearson, T. (2018). A review of Ofsted's test of the reliability of short inspections. Accessed 10/01/2022 from https://www.researchgate.net/publication/327894743_A_review_of_Ofsted's_test_of_the_reliability_of_short_inspections
Powell, D., & Lorenz, R. (2019). The Effect of Team Size on the Performance of Continuous Improvement Teams: Is Seven Really the Magic Number? In Farhad Ameri, Kathryn E. Stecke, Gregor von Cieminski and Dimitris Kiritsis (editors) *Advances in Production Management Systems. Production Management for the Factory of the Future*. Springer International Publishing
Richards, C. (2012). Ofsted Inspection Inspected: an examination of the 2012 framework for school inspection and its accompanying evaluation schedule. *FORUM, 54*(2), 247–272. https://doi.org/10.2304/forum.2012.54.2.247
Richardson, H. (2015). Ofsted purges 1,200 'not Good enough' inspectors. *BBC News Website 19th June 2015.* Accessed 10/01/2022 from https://www.bbc.co.uk/news/education-33198707
Richmond, T. (2019). *Requires Improvement: A new role for Ofsted and school inspections*. EDSK. Accessed 17/02/2022 from https://www.edsk.org/wp-content/uploads/2019/04/Requires-Improvement.pdf
Shapiro, J. R., Williams, A. M., & Hambarchyan, M. (2013). Are all interventions created equal? A multi-threat approach to tailoring stereotype threat interventions. *Journal of Personality and Social Psychology*, *104*(2), 277–288. https://doi.org/10.1037/a0030461
Spielman, A. (2017). HMCI's commentary: new research into short school inspections. Accessed 17/02/2022 from https://www.gov.uk/government/speeches/hmcis-monthly-commentary-march-2017
Steffensmeier, D., & Hebert, C. (1999). Women and men policymakers: does the judge's gender affect the sentencing of criminal defendants? *Social Forces, 77*(3), 1163–1196. https://doi.org/10.2307/3005975
Süß, S., & Kleiner, M. (2007). The psychological relationship between companies and freelancers: an empirical study of the commitment and the work-related expectations of freelancers. *Management Revue, 18*(3), 251–270. http://doi.org/10.5771/0935-9915-2007-3-251
Verbree, A. R., Hornstra, L., Maas, L., & Wijngaards-de Meij, L. (2022). Conscientiousness as a predictor of the gender gap in academic achievement. *Research in Higher Education*. https://doi.org/10.1007/s11162-022-09716-5

Villanueva-Moya, L., & Expósito, F. (2021). Gender differences in decision-making: the effects of gender stereotype threat moderated by sensitivity to punishment and fear of negative evaluation. *Journal of Behavioral Decision Making*, *34*(5), 706–717. https://doi.org/10.1002/bdm.2239

Conflict of interest statement: John Jerrim is currently a part-time specialist advisor to Ofsted on academic research on secondment from UCL. This paper is part of his independent research conducted as an academic at UCL.

Author contribution statement: Bokhove, Jerrim and Sims are joint first authors. They made an equal contribution to the conceptualisation of the research and the methodology used. Jerrim and Bokhove have led on extracting inspector names and linking this with publicly available data on Ofsted inspection outcomes. Jerrim led the writing of the introduction and data sections. The methodology, analysis, results and concluding sections were co-produced jointly by the three authors.

Table 1. The frequency and consequences of Ofsted inspections

Inspection type	Outcome	Consequence	Date rule introduced
Full (S5) inspection	Outstanding	Exempt from routine inspection.	May 2012 – November 2020
	Good	Short inspection within next 4 years.	September 2015 – present
	RI	Another full inspection within 30 months. Two consecutive RI judgements leads to monitoring visits.	
	Inadequate	Academy order issued (maintained schools). Monitoring visits and a full inspection within 30 months (academy).	
Short (S8) inspection	Previous grade maintained	Another short inspection within 4 years	September 2015 – present
	Convert to full inspection	Full inspection conducted within 48 hours	September 2015 – present
	S5 next	A full inspection to be conducted within the next 1 to 2 years	September 2018 – present

Table 2. Conversions from short to full inspections pre/post January 2018. Primary schools.

	Sep15 – Dec17		Jan18 – Aug19	
	N	%	N	%
Did not convert				
Remains Good (did not convert)	3,779	80%	2,840	78%
Converted to new grade				
Convert to Inadequate	110	2%	14	0.4%
Convert to Outstanding	270	6%	0	0%
Convert to RI	543	12%	4	0.1%
Recommend full inspection next				
Full inspection next (concerns)	-	-	376	10%
Full inspection next (progress)	-	-	393	11%
Total	**4,702**	**100%**	**3,627**	**100%**

Table 3. Descriptive statistics for the distribution of inspector characteristics

	Primary		Secondary	
	Short	**Not short**	**Short**	**Not short**
Lead inspector contract				
HMI	59%	20%	60%	45%
OI	41%	80%	40%	55%
Lead inspector gender				
Female	54%	48%	44%	43%
Male	45%	51%	55%	56%
unknown	0%	0%	0%	0%
Primary / secondary specialism				
Primary inspections only	65%	73%	0%	0%
70-99% primary	19%	17%	7%	13%
30-69% primary	14%	9%	38%	40%
Secondary inspections only	1%	1%	55%	47%
Inspection outside home region				
Yes	2%	16%	3%	15%
No	90%	67%	83%	62%
Not available	7%	17%	14%	23%
Academic year				
2011/12	0%	19%	0%	15%
2012/13	0%	23%	0%	21%
2013/14	0%	19%	0%	17%
2014/15	0%	15%	0%	14%
2015/16	12%	6%	18%	7%
2016/17	33%	5%	36%	7%
2017/18	34%	6%	32%	9%
2018/19	21%	7%	15%	9%
Previous inspections led				
mean	33	29	19	19
std	30	28	17	20
min	1	1	1	1
25th percentile	11	8	6	5
50th percentile	25	20	14	12
75th percentile	43	42	26	26
max	186	182	103	161
Team size				
1 inspector	83%	28%	12%	7%
2 inspectors	6%	35%	56%	7%
3 inspectors	6%	33%	8%	26%
4 inspectors	4%	4%	9%	44%
5	1%	0%	15%	15%
N	**8329**	**21521**	**1199**	**4747**

Table 4. The percentage of the variation in inspection outcomes that occurs between different inspectors

(a) Primary

	Primary	
	Unconditional	Conditional
Ordinal	9.6%	9.4%
Outstanding	9.6%	8.3%
Good	6.3%	5.1%
RI	7.2%	6.9%
Inadequate	17.5%	16.0%
Short inspection	12.2%	10.8%

(b) Secondary

	Secondary	
	Unconditional	Conditional
Ordinal	7.2%	5.1%
Outstanding	9.2%	12.0%
Good	4.9%	2.7%
RI	4.0%	2.2%
Inadequate	11.5%	10.2%
Short inspection	5.0%	0.4%

Notes: Figures refer to the percent of the variation in inspection outcomes that occurs between different inspectors. Estimates based upon multi-level (random effects) ordinal or binary logistic regression models, with inspections being nested within inspectors. Unconditional estimates refer to results from an empty model with no controls. Conditional estimates include controls for percent of pupils eligible for Free School Meals, Ofsted region, previous Ofsted inspection rating, inspection type, school gender composition (secondary only) and school performance measures (average Key Stage 2 maths and English scores for primary schools and average Key Stage 4 grades and progress measures for secondary schools). Primary estimates based upon 22,761 inspections conducted by 996 inspectors (other than short inspections, which is based upon 8,329 inspections conducted by 565 inspectors). Secondary estimates based upon 5,024 inspections conducted by 586 inspectors (other than short inspections, which is based upon 1,199 inspections conducted by 253 inspectors). Analysis based upon all inspections conducted between the 2011/12 and 2018/19 academic years.

Table 5. Crosstabulation between the gender of the lead inspector and overall effectiveness judgements

	Primary			Secondary		
	Female	**Male**	**Difference**	**Female**	**Male**	**Difference**
Outstanding	7.8%	8.2%	0.4%	10.9%	10.1%	-0.9%
Good	55.9%	58.7%	2.9%	45.4%	44.9%	-0.5%
Requires Improvement	30.5%	28.6%	-1.9%	34.6%	34.6%	-0.1%
Inadequate	5.9%	4.5%	-1.4%	9.1%	10.5%	1.4%
# of inspections	**11,056**	**11,698**		**2,188**	**2,813**	

Notes: Figures refer to column percentages.

Table 6. Differences in inspection assignments by gender and contract status of the lead inspector

	Gender		HMI	
	Female	Male	OI	HMI
Inspection type				
S5	68%	70%	74%	49%
RI reinspection	20%	18%	18%	22%
Academy first S5	5%	4%	3%	10%
S8 deemed S5	4%	4%	4%	7%
Serious weakness inspection	1%	1%	1%	4%
Exempt school inspection	2%	2%	0%	8%
S8 no formal designation	0%	0%	0%	1%
Missing	0%	0%	0%	0%
Prior inspection rating				
Outstanding	8%	8%	7%	13%
Good	41%	40%	42%	37%
RI	43%	45%	47%	32%
Inadequate	4%	4%	2%	14%
Missing	3%	3%	3%	5%
FSM quintile				
Q1 (Low FSM)	16%	17%	17%	15%
Q2	19%	20%	20%	18%
Q3	21%	21%	21%	20%
Q4	22%	22%	21%	24%
Q5 (High FSM)	22%	20%	20%	23%
Missing	0%	0%	0%	0%
School absence quintile				
Q1 (low absences)	20%	21%	21%	19%
Q2	23%	22%	23%	21%
Q3	23%	23%	23%	24%
Q4	21%	21%	21%	22%
Q5 (high absences)	13%	13%	13%	14%
Missing	0%	0%	0%	0%
KS2 English quintile				
Q1 (low achievement)	24%	24%	23%	31%
Q2	21%	20%	21%	20%
Q3	17%	17%	18%	14%
Q4	17%	16%	17%	14%
Q5 (high achievement)	12%	13%	13%	9%
Missing	9%	10%	10%	10%
KS2 maths quintile				
Q1 (low achievement)	24%	23%	22%	31%
Q2	19%	20%	20%	18%
Q3	20%	18%	19%	20%
Q4	15%	15%	16%	12%
Q5 (high achievement)	12%	13%	13%	10%
Missing	9%	10%	10%	10%

Table 7. Ordinal regression model estimates of the link between inspector gender and inspection outcomes. Primary school results.

	M0		M1		M2		M3		M4		M5		M6	
	OR	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**
Female inspector (ref: Male)	0.87*	-2.60	0.86*	-2.87	0.85*	-3.05	0.84*	-3.18	0.83*	-3.39	0.84*	-3.22	0.84*	-3.18
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector an HMI	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Inspection experience	-		-		-		-		-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Odds-ratios below one indicate that being inspected by a female lead inspector is associated with a worse inspection outcome. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. M0-M3 based upon 22,754 inspections conducted by 983 inspectors. M4-M6 based upon 21,366 inspections conducted by 983 inspectors. Standard errors have been clustered at the inspector level.

Table 8. The link between inspector gender and primary school inspection outcomes. Sub-judgements.

	N	Odds ratio	T-stat
Behaviour	14863	0.84*	-2.48
Development	6503	0.81*	-3.02
Leadership & Management	21366	0.83*	-3.44
Outcomes	7588	0.81*	-3.21
Quality	6503	0.84*	-2.41
Teaching	14863	0.86*	-2.19
Overall effectiveness	21366	0.84*	-3.18

Notes: See notes to Table 7 for further details. Number of observations differs due to sub-domains changing over time. Estimates based model specification M6, which controls for percentage of children eligible for FSM, school religion, school gender composition, Ofsted region, inspection type, prior school inspection rating, school performance data, school absences, percentage of pupils with SEN, percentage of pupils with EAL, whether the inspector is an HMI, inspectors amount of inspection experience, inspector phase specialism (primary versus secondary) and whether the inspector is inspecting in their home region.

Table 9. The association between lead inspector gender and short inspection outcomes

(a) Primary

	Time-period	N	Odds ratio	Confidence interval
S5 next due to concerns or conversion with a subsequent downgrade in judgement	September 2015 – August 2019	8,302	0.81	0.66 – 0.99
Conversion leading to RI / Inadequate	September 2015 – December 2017	4697	0.85	0.67 – 1.08
Conversion or S5 recommended next due to concerns	January 2018 – August 2019	3605	0.75	0.57 – 0.97

(b) Secondary

	Time-period	N	Odds ratio	Confidence interval
S5 next due to concerns or conversion with a subsequent downgrade in judgement	September 2015 – August 2019	1,184	0.93	0.65-1.29
Conversion leading to RI / Inadequate	September 2015 – December 2017	753	1.19	0.75-1.75
Conversion or S5 recommended next due to concerns	January 2018 – August 2019	431	0.84	0.45-1.39

Notes: Odds ratios below one indicates that male lead inspectors are less likely to convert or recommend an S5 inspection next than their female counterparts. Estimates based upon model M6 (see Table 7 for further details).

Table 10. The unconditional association between inspector contract status and Ofsted inspection judgements

	Primary			Secondary		
	OI	**HMI**	**Difference**	**OI**	**HMI**	**Difference**
Outstanding	7.7%	9.0%	-1.3%	10.5%	10.3%	-0.2%
Good	60.3%	47.0%	-13.3%	47.8%	42.2%	-5.6%
Requires Improvement	27.8%	35.4%	7.7%	33.9%	35.4%	1.6%
Inadequate	4.2%	8.6%	4.4%	7.8%	12.1%	4.3%
# of inspections	**17,622**	**5,139**		**2,654**	**2,370**	

Notes: Difference column refers to percentage for HMI minus percentage for OI. Analysis based upon data from 986 primary and 586 secondary lead inspectors. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years.

Table 11. Ordinal regression model estimates of the link between contract status and inspection outcomes. Primary school results.

	M0		M1		M2		M3		M4		M5		M6	
	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat
HMI (ref: OI)	1.54**	7.23	1.53**	7.55	1.49**	6.94	1.43**	6.12	1.43**	6.07	1.42**	5.95	1.45**	6.21
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector gender	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Inspection experience	-		-		-		-		-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * and ** Indicates that the estimates are statistically significant at the ten and five percent levels. Odds-ratios above one indicates that being inspected by an HMI is associated with a worse inspection outcome. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. M0-M3 based upon 22,761 inspections conducted by 986 inspectors. M4-M6 based upon 21,372 inspections conducted by 986 inspectors. Standard errors have been clustered at the inspector level.

Table 12. Ordinal regression model estimates of the link between contract status and inspection outcomes. Secondary school results.

	M0		M1		M2		M3		M4		M5		M6	
	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat
HMI (ref: OI)	1.26**	3.05	1.20**	2.72	1.13*	1.69	1.18**	2.14	1.18**	2.09	1.18**	2.03	1.32**	3.49
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector gender	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Inspection experience	-		-		-		-		-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. ** and * Indicates that the estimates are statistically significant at the five and ten percent levels. Odds-ratios above one indicates that being inspected by an HMI is associated with a worse inspection outcome. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. M0-M3 based upon 5,024 inspections conducted by 586 inspectors. M4-M6 based upon 4,899 inspections conducted by 564 inspectors. Standard errors have been clustered at the inspector level.

Table 13. The link between inspector contract status and inspection outcomes. Sub-judgements.

	Primary			Secondary		
	N	OR	T-stat	N	OR	T-stat
Behaviour	14868	1.59*	4.69	3053	1.66*	4.40
Development	6504	1.15	1.87	1846	0.97	-0.20
Leadership & Management	21372	1.38*	5.26	4899	1.31*	3.10
Outcomes	7589	1.33*	4.00	2028	1.10	0.79
Quality	6504	1.23*	2.63	1846	1.00	0.00
Teaching	14868	1.56*	4.93	3053	1.53*	4.24
Overall effectiveness	**21372**	**1.45***	**6.21**	4899	**1.32***	**3.49**

Notes: See notes to Tables 11 and 12 for further details. Number of observations differs due to sub-domains changing over time. Estimates based model specification M6 (see Tables 11 and 12 for further details). * indicates statistical significance at the 5% level.

Table 14. The unconditional association between inspector contract status and short inspection outcomes

	Primary		Secondary	
	OI	**HMI**	**OI**	**HMI**
Existing grade retained	90%	86%	82%	78%
Conversion with downgrade or S5 next due to concerns	10%	14%	18%	22%
# of inspections	**3,369**	**4,960**	**469**	**730**

Table 15. The association between lead inspector contract status and a negative outcome from a short inspection

(a) Primary

	Time-period	N	Odds ratio	Confidence interval
S5 next due to concerns or conversion with a subsequent downgrade in judgement	September 2015 - August 2019	8,329	1.44	1.18-1.76
Conversion leading to RI / Inadequate	September 2015 - December 2017	4702	1.27	1.00-1.62
Conversion or S5 recommended next due to concerns	January 2018 - August 2019	3627	1.63	1.25-2.14

(b) Secondary

	Time-period	N	Odds ratio	Confidence interval
S5 next due to concerns or conversion with a subsequent downgrade in judgement	September 2015 - August 2019	1,184	1.40	0.88-2.22
Conversion leading to RI / Inadequate	September 2015 - December 2017	753	1.56	0.89-2.71
Conversion or S5 recommended next due to concerns	January 2018 - August 2019	431	1.41	0.66-3.01

Notes: Odds ratios above one indicates that short inspection led by an HMI more likely to lead to a negative outcome for the school than a short inspection led by an OI. Estimated based upon model M6 – see Table 12 for further details.

Table 16. Ordinal regression model estimates of the link between inspector experience and overall effectiveness judgements.

(a) Primary schools

	Q2		Q3		Q4		Q5		
Model	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**N**
0	0.94	-0.63	0.99	-0.09	1.13	1.40	0.89	-1.26	6523
1	1.00	-0.03	1.06	0.69	1.15	1.55	0.93	-0.70	6523
2	1.10	1.00	1.14	1.46	1.21*	2.12	0.99	-0.09	6523
3	1.00	0.00	1.04	0.39	1.12	1.26	0.94	-0.72	6505
4	0.97	-0.28	1.02	0.24	1.09	0.95	0.87	-1.44	6505
5	0.95	-0.50	1.02	0.24	1.08	0.84	0.87	-1.43	6505
6	0.96	-0.47	1.03	0.31	1.09	0.95	0.89	-1.22	6505
7	0.95	-0.56	1.01	0.08	1.07	0.70	0.92	-0.89	6505
8	0.95	-0.48	1.02	0.22	1.09	0.91	0.95	-0.51	6505
9	0.94	-0.60	1.00	0.03	1.07	0.69	0.93	-0.64	6505

(b) Secondary schools

	Q2		Q3		Q4		Q5		
Model	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**N**
0	0.90	-0.75	0.95	-0.41	0.94	-0.45	1.11	0.68	1884
1	0.92	-0.56	0.97	-0.26	0.95	-0.36	1.14	0.77	1884
2	0.84	-1.22	0.96	-0.34	0.94	-0.43	1.04	0.27	1884
3	0.80	-1.53	0.91	-0.75	0.93	-0.48	1.03	0.18	1872
4	0.80	-1.49	0.75*	-2.02	0.85	-1.02	0.89	-0.59	1847
5	0.76	-1.83	0.73*	-2.15	0.83	-1.17	0.90	-0.51	1847
6	0.76	-1.83	0.74*	-2.06	0.83	-1.15	0.90	-0.51	1847
7	0.76	-1.84	0.74*	-2.07	0.83	-1.17	0.90	-0.50	1847
8	0.76	-1.77	0.75	-1.95	0.85	-0.89	0.94	-0.32	1847
9	0.72	-1.95	0.71*	-2.13	0.82	-1.08	0.89	-0.53	1847

Notes: Sample of inspections from September 2015 to August 2019. Experience measured as total number of inspections conducted since September 2011. Low experience (Q1) is the reference group. OR refers to the estimated odds ratio. * indicates statistical significance at the 5 level. M0 has no controls. M1 adds a control for academic year. M2 controls for school religion, gender, FSM and Ofsted region. M3 adds controls for prior inspection outcome and inspection type. M4 adds school performance data. M5 controls for school absence, EAL and SEN. M6 adds inspector gender. M7 adds inspector contract status (HMI / OI). M8 adds whether inspector phase specialism (primary/secondary). M9 controls for whether inspection was conducted outside the inspector's home region.

Table 17. Cross-tabulation for whether an inspection took place outside of the lead inspector's home region and inspection outcomes

(a) Primary

	Inspection outside of home region	
	No %	**Yes %**
Overall effectiveness		
Outstanding	8	7
Good	57	59
Requires Improvement	30	28
Inadequate	5	5
N	**15,925**	**3,347**
S5 next due to concerns or conversion with subsequent downgrade		
No	87	86
Yes	13	14
N	**7,627**	**183**

(b) Secondary

	Inspection outside of home region	
	No %	**Yes %**
Overall effectiveness		
Outstanding	9	15
Good	45	44
Requires Improvement	36	34
Inadequate	11	7
N	**3,161**	**735**
S5 next due to concerns or conversion with subsequent downgrade		
No	80	72
Yes	20	28
N	**988**	**43**

Notes: Figures refer to column percentages. Lower panel captures whether the short inspection was converted to a full inspection with a subsequent downgrade or a recommendation was made for an S5 inspection to be conducted next.

Table 18. Ordinal logistic regression model estimates of the link between whether the inspection was within the lead inspector's home region and overall effectiveness judgements. Results for primary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**
Outside region	0.99	-0.20	1.02	0.42	1.06	1.24	1.08	1.61	1.09	1.78	1.11*	2.03	1.13*	2.50	1.13*	2.53	1.13*	2.50
N	**19272**		**19272**		**19254**		**19254**		**18122**		**18122**		**18122**		**18122**		**18122**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Phase specialism	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that inspections conducted outside of the lead inspector's home region have worse inspection outcomes.

Table 19. Ordinal logistic regression model estimates of the link between whether the inspection was within the lead inspector's home region and overall effectiveness judgements. Results for secondary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**
Outside region	0.73**	-3.44	0.92	-0.93	0.96	-0.46	0.97	-0.37	0.96	-0.39	0.96	-0.38	1.00	-0.04	1.00	0.00	0.99	-0.14
N	**3,896**		**3,896**		**3,884**		**3,835**		**3,793**		**3,793**		**3,793**		**3,793**		**3,793**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Phase specialism	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that inspections conducted outside of the lead inspector's home region have worse inspection outcomes.

Table 20. Cross-tabulation between the percent of primary school inspections an inspector conducts throughout their inspection career and Ofsted judgements

(a) Primary

	Phase specialism		
	30-69% primary	**70-99% primary**	**Primary only**
Overall effectiveness			
Outstanding	11	9	7
Good	53	56	58
Requires Improvement	30	29	30
Inadequate	6	5	5
N	**1,912**	**4,880**	**15,871**
S5 next due to concerns or conversion with downgrade			
No	87	85	88
Yes	13	15	12
N	**1,062**	**1,833**	**5,378**

(b) Secondary

	Phase specialism		
	30-69% primary	**70-99% primary**	**Secondary only**
Overall effectiveness			
Outstanding	10	9	11
Good	44	46	46
Requires Improvement	35	37	33
Inadequate	11	8	10
N	**1,976**	**740**	**2,308**
S5 next due to concerns or conversion with downgrade			
No	77	75	82
Yes	23	25	18
N	**425**	**107**	**667**

Notes: Figures refer to column percentages. Lower panel captures whether the short inspection was converted to a full inspection with a subsequent downgrade or a recommendation was made for an S5 inspection to be conducted next.

Table 21. Ordinal logistic regression model estimates of the link between inspector phase specialism and overall effectiveness judgements. Results for primary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**
Primary only (Ref.)																		
30-69% primary	0.97	-0.31	0.96	-0.50	0.91	-1.17	0.89	-1.49	0.87	-1.73	0.89	-1.44	0.85*	-2.04	0.86	-1.92	0.86	-1.91
70-99% primary	0.96	-0.54	0.97	-0.41	0.96	-0.65	0.95	-0.76	0.94	-0.88	0.95	-0.85	0.95	-0.86	0.95	-0.83	0.95	-0.83
	22663		**22663**		**22645**		**22645**		**21277**		**21277**		**21277**		**21277**		**21277**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Phase specialism	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that those inspectors who have ever conducted secondary school inspections have worse inspection outcomes.

Table 22. Ordinal logistic regression model estimates of the link between inspector phase specialism and overall effectiveness judgements. Results for secondary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**
Secondary only (Ref)																		
30-69% primary	1.12	1.33	0.96	-0.55	0.92	-1.09	0.92	-1.03	0.92	-1.02	0.92	-1.06	0.90	-1.36	0.92	-1.06	0.92	-1.06
70-99% primary	1.08	0.63	1.01	0.14	1.00	0.00	0.98	-0.19	0.99	-0.05	1.00	-0.02	1.01	0.11	1.03	0.24	1.03	0.21
	5024		**5024**		**5012**		**4958**		**4899**		**4899**		**4899**		**4899**		**4899**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Phase specialism	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that those inspectors who have ever conducted secondary school inspections have worse inspection outcomes.

Table 23. Ordinal logistic regression model estimates of the link between inspector phase specialism and a negative outcome from the short inspection. Results for primary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat	OR	T-Stat
Primary only (Reference)																		
30-69% primary	1.09	0.71	1.07	0.60	1.07	0.60	1.08	0.55	1.08	0.57	1.14	0.96	1.03	0.23	1.02	0.12	1.00	0.01
70-99% primary	1.38*	2.20	1.30	1.88	1.30	1.88	1.30	1.85	1.30	1.84	1.32	1.95	1.29	1.88	1.27	1.71	1.24	1.58
	8273		**8273**		**8273**		**8273**		**8273**		**8273**		**8273**		**8273**		**8273**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Outside region	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that those inspectors who have ever conducted secondary school inspections are more likely to convert to full inspection leading to a downgrade or recommend an S5 inspection next due to concerns.

Table 24. Ordinal logistic regression model estimates of the link between inspector phase specialism and a negative outcome from the short inspection. Results for secondary schools.

	M0		M1		M2		M3		M4		M5		M6		M7		M8	
	OR	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**	**OR**	**T-Stat**
Secondary only (Ref)																		
30-69% primary	1.32	1.66	1.23	1.23	1.23	1.23	1.45	2.00	1.49	2.16	1.51	2.18	1.37	1.55	1.55	2.10	1.54*	2.07
70-99% primary	1.51	1.78	1.86*	2.68	1.86*	2.68	1.83*	2.62	1.75*	2.31	1.75*	2.32	1.69*	2.15	1.96*	2.62	2.00*	2.70
	1199		**1199**		**1199**		**1184**		**1184**		**1184**		**1184**		**1184**		**1184**	
Controls																		
School FSM	-		Y		Y		Y		Y		Y		Y		Y		Y	
School background	-		Y		Y		Y		Y		Y		Y		Y		Y	
Region	-		Y		Y		Y		Y		Y		Y		Y		Y	
Previous rating	-		-		Y		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y		Y		Y	
School performance	-		-		-		Y		Y		Y		Y		Y		Y	
% EAL	-		-		-		-		Y		Y		Y		Y		Y	
%SEN	-		-		-		-		Y		Y		Y		Y		Y	
Male inspector	-		-		-		-		-		Y		Y		Y		Y	
Inspector an HMI	-		-		-		-		-		-		Y		Y		Y	
Outside region	-		-		-		-		-		-		-		Y		Y	
Post 2018	-		-		-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates that those inspectors who have ever conducted secondary school inspections are more likely to convert to full inspection leading to a downgrade or recommend an S5 inspection next due to concerns.

Table 25. Predicted probability of a negative outcome from the short inspection by inspector phase specialism.

	Phase specialism			
	30-69% primary	70-99% primary	Primary only	Secondary only
Ofsted Phase				
Primary	12%	14%	12%	-
Secondary	22%	26%	-	19%

Notes: Model controls for percentage of pupils eligible for free school meals, Ofsted region, prior inspection rating, inspection type, school performance measures, school absences, percentage of pupils who have English as an Additional Language, gender of the inspector, contract status of the inspector (OI or HMI) and whether the short inspection was conducted before or after January 2018.

Table 26. Cross-tabulation between inspection team size and inspection outcomes

(a) Primary

	Team size				
	1	**2**	**3**	**4**	**5**
Overall effectiveness					
Outstanding	9%	7%	8%	14%	26%
Good	61%	59%	55%	46%	30%
Requires Improvement	27%	29%	31%	33%	41%
Inadequate	3%	5%	7%	8%	3%
N	**5,546**	**7,184**	**7,158**	**1,093**	**150**
Short conversion with downgrade (Sep 15 - Dec 17)					
No	96%	73%	48%	54%	54%
Yes	4%	27%	52%	46%	46%
N	**3,458**	**241**	**444**	**341**	**103**
S5 next due to concerns or conversion with downgrade (Jan 18 - Aug 19)					
No	89%	90%	-	-	-
Yes	11%	10%	-	-	-
N	**3,279**	**206**	**-**	**-**	**-**

(b) Secondary

	Team size				
	1	**2**	**3**	**4**	**5**
Overall effectiveness					
Outstanding	22%	9%	7%	10%	15%
Good	43%	47%	46%	44%	46%
Requires Improvement	30%	36%	35%	35%	32%
Inadequate	4%	9%	12%	11%	7%
N	**233**	**273**	**1,148**	**2,072**	**889**
Short conversion with downgrade (Sep 15 - Dec 17)					
No	100%	99%	40%	52%	52%
Yes	0%	1%	60%	48%	48%
N	**74**	**353**	**40**	**95**	**168**
S5 next due to concerns or conversion with downgrade (Jan 18 - Aug 19)					
No	85%	79%	69%	-	-
Yes	15%	21%	31%	-	-
N	**71**	**307**	**45**	**-**	**-**

Notes: Figures refer to column percentages. Lower panel captures whether there was a negative outcome from the short inspection (conversion with a downgrade in Overall Effectiveness rating or recommendation of S5 next due to concerns).

Table 27. Ordinal logistic regression model estimates of the link between inspection team size and Overall Effectiveness judgements. Results for primary schools.

	M0		M1		M2		M3		M4		M5		M6	
	OR	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**
Team size (Ref: 1 inspector)														
2 inspectors	1.24*	5.93	1.15*	3.64	1.20*	4.88	1.18*	4.25	1.21*	4.64	1.21*	4.67	1.25*	5.29
3 inspectors	1.37*	8.13	1.25*	5.67	1.32*	7.02	1.18*	3.91	1.22*	4.63	1.23*	4.73	1.26*	5.23
4 inspectors	1.33*	3.43	1.33*	3.47	1.21*	2.34	0.98	-0.19	1.03	0.35	1.03	0.33	1.05	0.56
5 inspectors	0.99	-0.06	1.07	0.27	0.75	-1.19	0.61*	-2.04	0.66	-1.74	0.64	-1.87	0.68	-1.60
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector gender	-		-		-		-		-		Y		Y	
Inspector HMI	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Academic year	-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates a lower Overall Effectiveness judgement is reached than for the reference group (one inspector).

Table 28. Ordinal logistic regression model estimates of the link between inspection team size and Overall Effectiveness judgements. Results for secondary schools.

	M0		M1		M2		M3		M4		M5		M6	
	OR	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**
Team size (Ref: 4 inspectors)														
1 inspector	0.50*	-4.60	0.59*	-3.55	0.52*	-4.43	0.46*	-4.83	0.44*	-5.07	0.41*	-5.51	0.43*	-4.92
2 inspectors	0.97	-0.25	0.91	-0.80	0.89	-1.01	0.91	-0.78	0.87	-1.20	0.86	-1.32	0.90	-0.93
3 inspectors	1.14*	1.98	1.10	1.44	1.05	0.75	1.08	0.97	1.08	0.99	1.08	0.90	1.08	0.93
5 inspectors	0.71*	-4.19	0.82*	-2.44	0.79*	-2.72	0.84*	-2.05	0.84*	-2.01	0.83*	-2.04	0.83*	-2.10
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector gender	-		-		-		-		-		Y		Y	
Inspector HMI	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Academic year	-		-		-		-		-		-		Y	

Notes: * indicates statistical significance at the 5% level. Odds ratios above one indicates a lower Overall Effectiveness judgement is reached than for the reference group (four inspectors).

Table 29. Predicted distribution of primary school inspection outcomes for two hypothetical inspectors.

(a) Multi-nominal logistic regression estimates

	Inspector A	Inspector B	Risk ratio (A/B)
Overall effectiveness			
Outstanding	6.0%	7.8%	0.77
Good	45.4%	60.1%	0.76
Requires Improvement	35.3%	28.7%	1.23
Inadequate	13.3%	3.4%	3.94
Short inspection			
Conversion with downgrade or S5 next due to concerns. (Jan18 - Aug19)	15.5%	9.7%	1.60
Inspector characteristics			
Team size	2 inspectors	1 inspector	
Contract status	HMI	OI	
Gender	Female	Male	

(b) Ordinal logistic regression estimates

	Inspector A	Inspector B	Risk ratio (A/B)
Overall effectiveness			
Outstanding	4.5%	9.0%	0.50
Good	48.0%	59.3%	0.81
Requires Improvement	38.4%	27.2%	1.41
Inadequate	9.1%	4.5%	2.03
Short inspection			
Conversion with downgrade or S5 next due to concerns. (Jan18 - Aug19)	15.5%	9.7%	1.60
Inspector characteristics			
Team size	2 inspectors	1 inspector	
Contract status	HMI	OI	
Gender	Female	Male	

Notes: Multinominal logistic estimates control for percent of pupils eligible for FSM, region, previous Ofsted inspection outcome, inspection type, Key Stage 2 maths and English scores, school absences, percent of pupils with English as an additional language and whether the inspection was conducted after 2018. Ordinal logistic regression models additionally control for school religion, school gender composition, Key Stage 1 scores and percent of pupils with special educational needs.

Appendix A. Inspection type of academic year. 2011/12 – 2018/19

Appendix Table A1. Cross-tabulation between inspection type by academic year. Primary inspections.

Inspection type	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19
Academy First Section 5	0	4	164	238	18	250	209	152
Exempt School Inspection	0	0	0	0	5	96	68	280
Maintained Academy and School Short inspection	0	0	0	0	1,091	2,903	3,034	1,906
Notice to Improve S5 Reinspection	101	131	0	0	0	0	0	0
Requires Improvement S5 Reinspection Visit 1	0	0	0	0	1,139	535	177	365
Requires Improvement S5 Reinspection Visit 2	0	0	0	0	96	217	122	156
Requires Improvement S5 Reinspection Visit 3	0	0	0	0	0	0	3	14
Requires Improvement monitoring Visit 1	0	0	0	0	275	259	234	83
Requires Improvement monitoring Visit 2	0	0	0	0	49	22	6	8
Requires Improvement monitoring Visit 3	0	0	0	0	2	3	0	0
S5 Inspection	4,370	5,588	3,407	1,924	9	97	888	855
S5 Requires Improvement 1st Re-Inspection	0	0	567	1,079	1	0	0	0
S5 Requires Improvement 2nd Re-Inspection	0	0	0	1	0	0	0	0
S5 Serious Weaknesses Re-Inspection	0	0	29	37	0	0	0	0
S8 Deemed S5	147	115	649	360	0	0	0	0
S8 No Formal Designation Visit	0	0	0	0	48	40	51	33
Schools into Special Measures Visit 1	0	0	0	0	50	43	61	52
Schools into Special Measures Visit 2	0	0	0	0	55	14	45	15
Schools into Special Measures Visit 3	0	0	0	0	58	7	28	16
Schools into Special Measures Visit 4	0	0	0	0	54	4	8	16
Schools into Special Measures Visit 5	0	0	0	0	33	5	5	5
Schools with Serious Weaknesses Visit 1	0	0	0	0	7	5	22	30
Schools with Serious Weaknesses Visit 2	0	0	0	0	15	2	4	3
Schools with Serious Weaknesses Visit 3	0	0	0	0	4	0	0	0
Section 8 Inspection due to Parental Complaint	0	0	0	0	22	14	10	4
Serious Weaknesses S5 Reinspection	0	0	0	0	19	3	3	13
Special Measures S5 Reinspection	18	8	10	17	14	6	7	22
Total	**4,636**	**5,846**	**4,826**	**3,656**	**3,064**	**4,525**	**4,985**	**4,028**

Appendix Table A2. Cross-tabulation between inspection type by academic year. Secondary inspections.

Inspection type	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19
Academy First Section 5	70	109	67	72	6	52	64	52
Exempt School Inspection	-	-	-	-	2	16	10	68
Maintained Academy and School Short inspection	-	-	-	-	243	470	418	193
Notice to Improve S5 Reinspection	38	52	1	-	-	-	-	-
Requires Improvement S5 Reinspection Visit 1	-	-	-	-	263	159	109	104
Requires Improvement S5 Reinspection Visit 2	-	-	-	-	36	79	52	54
Requires Improvement S5 Reinspection Visit 3	-	-	-	-	-	-	-	13
Requires Improvement monitoring Visit 1	-	-	-	-	119	89	89	69
Requires Improvement monitoring Visit 2	-	-	-	-	32	19	2	2
Requires Improvement monitoring Visit 3	-	-	-	-	4	-	-	-
S5 Inspection	784	1,152	645	302	11	73	232	194
S5 Requires Improvement 1st Re-Inspection	-	-	107	315	-	-	-	-
S5 Requires Improvement 2nd Re-Inspection	-	-	-	1	-	-	-	-
S5 Serious Weaknesses Re-Inspection	-	-	25	28	-	-	-	-
S8 Deemed S5	30	23	203	140	-	-	-	-
S8 No Formal Designation Visit	-	-	-	-	35	52	58	47
Schools into Special Measures Visit 1	-	-	-	-	35	38	60	34
Schools into Special Measures Visit 2	-	-	-	-	36	24	42	23
Schools into Special Measures Visit 3	-	-	-	-	57	19	25	14
Schools into Special Measures Visit 4	-	-	-	-	55	17	16	15
Schools into Special Measures Visit 5	-	-	-	-	53	15	7	7
Schools with Serious Weaknesses Visit 1	-	-	-	-	9	13	16	24
Schools with Serious Weaknesses Visit 2	-	-	-	-	11	2	3	3
Schools with Serious Weaknesses Visit 3	-	-	-	-	9	-	1	-
Section 8 Inspection due to Parental Complaint	-	-	-	-	25	13	16	9
Serious Weaknesses S5 Reinspection	-	-	-	-	22	8	8	14
Special Measures S5 Reinspection	4	1	-	12	26	11	8	17
Total	**926**	**1,337**	**1,048**	**870**	**1,089**	**1,169**	**1,236**	**956**

Appendix B. Sample selection

Primary inspections

Between September 2011 and August 2019 there were 35,566 primary inspections conducted, based upon the management information published on the Ofsted website[15]. We have extracted information on the lead inspector from 29,850 (84%) of these inspections from the "Watchsted" website. Of the remaining 5,716 inspections, we can access information on the inspectors involved in the inspection from 3,776 via our own scraping of the published Ofsted reports. Thus, when added together, we can observe information on the lead inspector from (29,850 + 3,776) / 35,566 = 94.5% of all primary inspections conducted between September 2011 and August 2019[16]. Moreover, of the 1,940 (5.5%) of primary inspections we have been unable to match, 1,489 are Requires Improvement monitoring visits or special measures/serious weakness visits. Importantly, most (1,664 – 86%) of the 1,940 unmatched did not lead to an overall effectiveness judgement. Together, this provides reassurance that we have managed to access the relevant information on the vast majority of primary inspections conducted over this period, that issues of missing / unlinked data are limited, and that our analytic sample is representative of the population of primary inspections conducted over this period.

Secondary inspections

Between September 2011 and August 2019 there were 8,631 secondary inspections conducted, based upon the management information published on the Ofsted website. We have extracted information on the lead inspector from 5,901 (68%) of these inspections from the "Watchsted" website. Of the remaining 2,730 inspections, we can access information on the inspectors involved in the inspection from 1,432 via our own scraping of the published Ofsted reports. Thus, when added together, we can observe information on the lead inspector from (5,901 + 1,432) / 8,631 = 85% of all secondary inspections conducted between September 2011 and August 2019. Moreover, of the 1,298 (15%) of secondary inspections we have been unable to

[15] This is based upon Excel files published by Ofsted at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/920755/Management_information_-_state-funded_schools_1_September_2015_to_31_August_2019.xlsx and https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/485634/Management_information_-_schools_-_1_Sept_2005_to_31_August_2015.xlsx

[16] If we focus only upon primary inspections that led to an overall effectiveness rating, we have been able to access 22,760 from a total of 26,360 via the Watchsted website. Of the remaining 3,600, we have managed to access information from a further 3,324 via our own scraping of the Ofsted reports. We have hence been able to access the relevant information for 99% of all primary inspections conducted between September 2011 and August 2019 that led to an overall effectiveness judgement.

match, 1,009 are Requires Improvement monitoring visits or special measures/serious weakness visits. Importantly, most (1,166 – 90%) of the 1,298 unmatched did not lead to an overall effectiveness judgement. Together, this provides reassurance that we have managed to access the relevant information on the vast majority of secondary inspections conducted over this period, that issues of missing / unlinked data are limited, and that our analytic sample is representative of the population of secondary inspections conducted over this period.

Robustness test to using an alternative sample

In the main body of the paper we present results based upon data we have extracted from the Watchsted database alone. However, as noted above, we have also performed our own scraping of inspector names from the published Ofsted inspection reports, which we can add onto the Watchsted database. Tables B1 (primary) and B2 (secondary) below provides a comparison of the results across these two analytic samples. The estimated odds ratios and associated t-statistics are very similar, regardless of which sample is used. In other words, the estimates reported in the main text appear robust to further extension of our analytic sample via adding in data from our own scrapping of inspector names into the Watchsted database.

Appendix Table B1. A comparison of ordinal logistic regression estimates across alternative sample selections. Estimates for primary schools.

	Main sample		Alternative sample	
	Odds ratio	**T-Stat**	**Odds ratio**	**T-Stat**
Gender (Ref: female)				
Male	0.86*	-2.83	0.85*	-3.32
Contract (Ref: OI)				
HMI	1.43*	6.24	1.41*	6.21
Outside home region (Ref: No)				
Yes	1.12*	2.22	1.11*	2.18
Phase specalism (Ref: primary only)				
30-69% primary	0.88	-1.71	0.96	-0.48
70-99% primary	0.96	-0.55	1.00	-0.01
Secondary only	1.27	1.21	0.75	-1.74
Experience (Ref: Q1)				
Q2	0.98	-0.52	1.00	0.09
Q3	0.86*	-3.07	0.86*	-2.99
Q4	0.97	-0.57	0.93	-1.32
Q5	0.88*	-1.99	0.88*	-1.98
Team size (Ref: 1 inspector)				
2 inspectors	1.20*	4.63	1.17*	4.41
3 inspectors	1.17*	3.94	1.15*	3.80
4 inspectors	0.96	-0.53	0.92	-1.04
5 inspectors	0.57*	-2.28	0.60*	-2.24
Inspections	22,743		25,936	
Inspectors	986		1,407	

Notes: Models include controls for percent of pupils eligible for FSM, Ofsted region, previous inspection rating, inspection type and Key Stage 2 English and mathematics test scores. * indicates statistical significance at the 5% level.

Appendix Table B2. A comparison of ordinal logistic regression estimates across alternative sample selections. Estimates for secondary schools.

	Main sample		Alternative sample	
	Odds ratio	**T-Stat**	**Odds ratio**	**T-Stat**
Gender (Ref: female)				
Male	1.08	1.08	1.05	0.76
Contract (Ref: OI)				
HMI	1.24*	2.72	1.18*	2.24
Outside home region (Ref: No)				
Yes	1.03	0.30	1.00	0.00
Phase specalism (Ref: secondary only)				
30-69% primary	0.92	-0.85	1.01	0.14
70-99% primary	0.96	-0.34	1.10	0.89
Experience (Ref: Q1)				
Q2	0.89	-1.39	0.88	-1.64
Q3	0.83*	-1.99	0.77*	-3.03
Q4	0.88	-1.23	0.79*	-2.18
Q5	0.94	-0.49	0.83	-1.47
Team size (Ref: 4 inspectors)				
1 inspector	0.37*	-6.34	0.51*	-5.12
2 inspectors	0.87	-0.97	0.93	-0.69
3 inspectors	1.11	1.27	1.02	0.22
5 inspectors	0.83*	-2.17	0.80*	-2.83
Inspections	4,899		6,191	
Inspectors	564		733	

Notes: Models include controls for percent of pupils eligible for FSM, religious denomination of the school, gender composition of school, Ofsted region, previous overall inspection rating, inspection type, Key Stage 2 scores of intake, percent of pupils achieving five A*-C grades, Key Stage 4 total points score, Progress 8 scores, percent of pupils absent, percent of pupils who speak English as an Additional Language, percent of pupils with special educational needs.

Appendix C. Manual checks of the data

To check the quality of the data, we have performed some "manual" checks, returning to the initially published Ofsted reports to cross-reference the data we have extracted against.

To begin, we conducted a power calculation to understand the sample size required from our manual checks to give us a reasonable degree of accuracy. These power calculations were conducted assuming that there would be around 90% agreement between the Watchsted data (plus our automated inspector name extraction where the Watchsted data is missing) and our manual approach. These power calculations revealed that a sample size of 150 would yield a standard error of 2.4 percentage points[17], and thus resulting in a confidence interval between 85% and 95%. We deemed this sufficient to understand the likely degree of measurement error within our data.

Two sets of random samples were drawn. The first random sample was 150 short inspections. The second was 150 not-short inspections (108 of these were an S5 inspection)[18]. For each of these 300 inspections, we attempted to find the relevant inspection report on the Ofsted website and manually recorded (a) the name of all inspectors (including the lead inspector) and (b) whether the lead inspector (or any other inspector) was an HMI. These are then used as a basis to check the quality of the full database we use in our analysis.

Non-short inspections

Of the 150 inspections in our initial random sample, the original inspection report was available from the Ofsted website on 138 (92%) occasions, for which we can manually check our data against. Of these, the name of the lead inspector matches on 134 (97%) of occasions (95% confidence interval spans 94% to 100%). Moreover, two of the four instances where the sources did not agree may be due to typos ("June Robinson" rather than "Jean Robinson" and "Christine Huard" rather than "Christine Howard"). The level of agreement for whether an HMI or OI led an inspection was also high (93% with a confidence interval spanning from 89% to 98%). In other words, the level of agreement is extremely high.

[17] This can be computed via the formula sqrt((p*(1-p))/n).
[18] 6 of the 108 were Academy first section 5.

Short inspection results

All 150 of the short inspections in our random sample were found and accessed from the Ofsted website. Of these, the lead inspector matched on 145 (97%) of occasions. This is again a very high level of agreement and is reassuring regarding the quality of the data available.

The level of agreement of whether an HMI or OI led the short inspection was somewhat lower at 130 (86%) out of the 150 (confidence interval spanning from 80% to 91%). Further investigations of the data suggest that this may be due to individual inspectors changing contracts type over time (i.e. moving from being an OI to an HMI, or vice-versa). As the Watchsted database only includes a fixed flag at the inspector level for whether the named inspector is an HMI or not, this time dimension to contract status will not be captured.

We hence also investigate the level of agreement (for whether an HMI was involved in the inspection or not) between our manually extracted random sample and our own automated extraction of inspector names (and HMI status). An important advantage of our own extraction of inspector names (and HMI status) is that it has been done at the individual inspection level – and hence captures potential changes in OI/HMI status of individual inspectors over time.

Of the 150 short inspections in our random sample, we have managed to perform our own manual extraction successfully on 145 occasions. Of these, there was agreement on 141 occasions (97%) as to whether an HMI was involved in the inspection (confidence interval 95% to 100%)[19]. Hence data from our own automated extraction of inspector names – and, in particular, whether an HMI was involved in the inspection – provides a useful additional source of information that can be further used to investigate the robustness of our results (most notably, differences between HMI and OIs in short inspection outcomes).

[19] When performing a similar analysis for "non-short" inspections, we get 99% agreement (confidence interval 98% to 100%) between our automated extraction of HMI involvement in the inspections and our manual coding such information from the inspection reports. This is based upon 128 of the random sample of 150 "non-short" inspections where data is available from across the two sources.

Appendix D. Alternative estimates using multi-level modelling (random effects)

In the main body of the paper we use ordinal logistic regression – with standard errors clustered by inspector – to examine the association between various inspector characteristics and school inspection outcomes. An alternative approach to taking account the "clustering" of inspections within lead inspectors would be to estimate a multilevel model (with inspections as the level 1 unit and lead inspectors as the second level). In this appendix, we explore the similarity of results under these two approaches, focusing on the results for Overall Effectiveness judgements.

Appendix Table D1 presents results from such a comparison of methodological approaches for primary schools, referring to a model that controls for percent of pupils eligible for FSM, Ofsted region, previous inspection rating, inspection type and Key Stage 2 English and mathematics test scores. All inspector character characteristics and included in this model simultaneously. Figures on the left-hand side are from a multilevel (random effects) ordinal logistic regression model, while those on the right are from an ordinal logistic regression model with standard errors clustered within inspectors. Overall, parameter estimates (presented as odds ratios) and the associated t-statistics are very similar across the two approaches. The substantive conclusions reached are thus robust

An analogous comparison across methodological approaches for secondary schools is presented in Appendix Table D2. The model used controls for gender composition of school, Ofsted region, previous overall inspection rating, inspection type, Key Stage 2 scores of intake, percent of pupils achieving five A*-C grades, Key Stage 4 total points score, Progress 8 scores, percent of pupils absent, percent of pupils who speak English as an Additional Language, percent of pupils with special educational needs. Again, the estimated odds-ratios and the associated t-statistics do not substantive differ across the two approaches.

Appendix D1. A comparison of estimates from multilevel ordinal logistic regressions to ordinal logistic regression with clustered standard errors. Primary school results.

	Multi-level model		Clustered SE	
	Odds ratio	**T-Stat**	**Odds ratio**	**T-Stat**
Gender (Ref: female)				
Male	0.88*	-2.60	0.86*	-2.83
Contract (Ref: OI)				
HMI	1.45*	6.35	1.43*	6.24
Outside home region (Ref: No)				
Yes	1.16*	3.40	1.12*	2.22
Phase specialism (Ref: primary only)				
30-69% primary	0.88	-1.61	0.88	-1.71
70-99% primary	0.94	-0.95	0.96	-0.55
Secondary only	1.32	1.07	1.27	1.21
Experience (Ref: Q1)				
Q2	0.96	-0.83	0.98	-0.52
Q3	0.84*	-3.72	0.86*	-3.07
Q4	0.97	-0.71	0.97	-0.57
Q5	0.83*	-3.53	0.88*	-1.99
Team size (Ref: 1 inspector)				
2 inspectors	1.21*	4.97	1.20*	4.63
3 inspectors	1.19*	4.27	1.17*	3.94
4 inspectors	0.98	-0.25	0.96	-0.53
5 inspectors	0.57*	-2.95	0.57*	-2.28
Inspections	22,743		22,743	
Inspectors	986		986	

Notes: Models include controls for percent of pupils eligible for FSM, Ofsted region, previous inspection rating, inspection type and Key Stage 2 English and mathematics test scores. * indicates statistical significance at the 5% level.

Appendix D2. A comparison of estimates from multilevel ordinal logistic regressions to ordinal logistic regression with clustered standard errors. Secondary school results.

	Multi-level model		Clustered SE	
	Odds ratio	**T-Stat**	**Odds ratio**	**T-Stat**
Gender (Ref: female)				
Male	1.08	1.08	1.09	1.15
Contract (Ref: OI)				
HMI	1.24*	2.72	1.25*	2.77
Outside home region (Ref: No)				
Yes	1.03	0.30	1.02	0.17
Phase specialism (Ref: secondary only)				
30-69% primary	0.92	-0.85	0.91	-1.03
70-99% primary	0.96	-0.34	1.01	0.08
Experience (Ref: Q1)				
Q2	0.89	-1.39	0.92	-1.00
Q3	0.83*	-1.99	0.84	-1.80
Q4	0.88	-1.23	0.93	-0.67
Q5	0.94	-0.49	1.03	0.24
Team size (Ref: 1 inspector)				
2 inspectors	0.37*	-6.34	0.38*	-5.81
3 inspectors	0.87	-0.97	0.85	-1.38
4 inspectors	1.11	1.27	1.08	0.93
5 inspectors	0.83*	-2.17	0.83*	-2.05
Inspections	4,899		4,899	
Inspectors	564		564	

Notes: Models include controls for percent of pupils eligible for FSM, religious denomination of the school, gender composition of school, Ofsted region, previous overall inspection rating, inspection type, Key Stage 2 scores of intake, percent of pupils achieving five A*-C grades, Key Stage 4 total points score, Progress 8 scores, percent of pupils absent, percent of pupils who speak English as an Additional Language, percent of pupils with special educational needs.

Appendix E. Sub-group ordinal logistic regression estimates for gender and contract status

Gender

Table E1. The link between inspector gender and primary school inspection outcomes. Ordinal regression estimates for sub-groups.

	Primary			Secondary		
	N	OR	T-stat	N	OR	T-stat
Academic year						
2011/12	3,587	0.79*	-2.45	583	0.73	-1.53
2012/13	5,092	0.80*	-2.43	1,034	1.09	0.56
2013/14	4,272	1.00	-0.02	800	1.10	0.64
2014/15	3,286	0.91	-1.07	693	1.25	1.20
2015/16	1,550	0.84	-1.39	406	1.33	1.15
2016/17	1,806	0.78*	-2.37	500	1.29	1.26
2017/18	1,528	0.92	-0.72	505	1.34	1.39
2018/19	1,621	0.84	-1.65	436	1.00	-0.01
Contract type						
Ofsted inspector (OI)	17,617	0.87*	-2.10	2,625	1.18	1.50
Her Majesty's Inspector (HMI)	5,126	0.83*	-2.36	2,333	0.99	-0.06
Inspection type						
S5 inspection	14,876	0.87*	-2.29	2,595	1.10	0.98
Ofsted region						
East Midlands	2,320	0.76*	-2.04	452	0.93	-0.27
East of England	2,788	0.97	-0.24	593	1.20	0.87
London	2,192	0.79*	-2.10	568	1.19	0.89
North East, Yorkshire and Humber	3,675	0.86	-1.14	746	1.02	0.08
North West	3,414	0.80*	-2.12	740	1.26	1.43
South East	3,454	0.86	-1.16	711	0.92	-0.49
South West	2,256	0.96	-0.27	510	1.06	0.20
West Midlands	2,644	0.98	-0.21	638	1.10	0.46

Notes: Estimates based upon ordered logistic regression models. The models control for percentage of children eligible for free school meals, Ofsted region, previous inspection rating, inspection type, school performance measures, whether the inspector is an HMI and total amount of inspection experience. Separate models have been estimated for each sub-group. * indicates statistical significance at the five percent level. Standard errors have been clustered at the inspector level.

Contract status

Table E2. The link between inspector contractual status and inspection outcomes. Ordinal regression estimates for sub-groups.

	Primary			Secondary		
	N	OR	T-stat	N	OR	T-stat
Academic year						
2011/12	3,587	1.93*	3.95	583	2.69*	4.78
2012/13	5,092	1.40*	2.74	1034	0.79	-1.35
2013/14	4,272	2.11*	6.08	800	2.26*	3.94
2014/15	3,286	1.53*	2.84	693	2.40*	3.56
2015/16	1,550	1.31	1.8	406	0.95	-0.20
2016/17	1,806	1.38*	2.69	500	0.91	-0.47
2017/18	1,528	1.50*	3.28	505	1.08	0.40
2018/19	1,621	1.09	0.64	436	0.96	-0.17
Inspection type						
S5 inspection	14,876	1.54*	5.84	2595	1.21*	1.97
Ofsted region						
East Midlands	2,320	1.46*	2.81	452	1.28	0.80
East of England	2,788	1.20	0.99	593	1.05	0.20
London	2,192	1.09	0.54	568	1.62*	2.70
North East, Yorkshire and Humber	3,675	1.49*	2.93	746	1.33	1.31
North West	3,414	1.49*	2.97	740	1.10	0.51
South East	3,454	1.09	0.58	711	1.00	0.01
South West	2,256	2.23*	4.93	510	1.43	1.37
West Midlands	2,644	1.74*	3.93	638	1.04	0.19

Notes: Estimates based upon ordered logistic regression models controlling for percentage of children eligible for free school meals, Ofsted region, previous inspection rating, inspection type, school performance measures, whether the inspector is male and total amount of inspection experience. Separate models have been estimated for each sub-group. * indicates statistical significance at the five percent level. Standard errors have been clustered at the inspector level.

Appendix F. Multinominal logistic regression model estimates

Gender

Table F1. The link between inspector gender and primary school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

	M0		M1		M2	
	OR	T-stat	OR	T-stat	OR	T-stat
Impact of female inspector (Ref: Good)						
Outstanding	1.00	-0.05	1.04	0.62	1.05	0.69
Requires Improvement	0.89	-2.11	0.88	-2.2	0.89	-2.03
Inadequate	0.73	-3.28	0.71	-3.44	0.73	-3.17
Inspection-level controls						
School % FSM	-		Y		Y	
Inspection type	-		Y		Y	
Prior Ofsted rating	-		Y		Y	
School performance data	-		Y		Y	
Inspector level controls						
Inspector an HMI	-		-		Y	

(b) Predicted probabilities

	M0		M1		M2	
	Female	Male	Female	Male	Female	Male
Impact of female inspector						
Outstanding	7.8%	8.2%	7.7%	8.2%	7.7%	8.2%
Good	55.9%	58.7%	55.6%	58.6%	56.1%	58.5%
Requires Improvement	30.5%	28.6%	30.4%	28.7%	30.4%	28.7%
Inadequate	5.9%	4.5%	5.8%	4.5%	5.8%	4.5%
Inspection-level controls						
School % FSM	-		Y		Y	
Inspection type	-		Y		Y	
Prior Ofsted rating	-		Y		Y	
School performance data	-		Y		Y	
Inspector level controls						
Inspector an HMI	-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level. Models based upon 22,736 inspections conducted by 983 inspectors.

Table F2. The link between inspector gender and secondary school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

	M0		M1		M2	
	OR	**T-stat**	**OR**	**T-stat**	**OR**	**T-stat**
Impact of female inspector (Ref: Good)						
Outstanding	0.93	-0.64	0.99	-0.10	0.99	-0.06
Requires Improvement	1.01	0.11	1.01	0.08	1.00	-0.03
Inadequate	1.16	1.15	1.16	1.19	1.11	0.86
Inspection-level controls						
School % FSM	-		Y		Y	
Inspection type	-		Y		Y	
Prior Ofsted rating	-		Y		Y	
School performance data	-		Y		Y	
Inspector level controls						
Inspector an HMI	-		-		Y	

(b) Predicted probabilities

	M0		M1		M2	
	Female	**Male**	**Female**	**Male**	**Female**	**Male**
Impact of female inspector						
Outstanding	10.9%	10.1%	10.5%	10.4%	10.5%	10.4%
Good	45.4%	44.9%	45.5%	44.9%	45.3%	45.0%
Requires Improvement	34.6%	34.6%	34.8%	34.3%	34.8%	34.3%
Inadequate	9.1%	10.5%	9.2%	10.4%	9.4%	10.2%
Inspection-level controls						
School % FSM	-		Y		Y	
Inspection type	-		Y		Y	
Prior Ofsted rating	-		Y		Y	
School performance data	-		Y		Y	
Inspector level controls						
Inspector an HMI	-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level. Models based upon 4,947 inspections conducted by 560 inspectors.

Contract status (OI versus HMI)

Table F3. The link between inspector gender and school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

	Primary				Secondary			
	M0		M1		M0		M1	
	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat
Impact of HMI (Ref: Good)								
Outstanding	1.51*	5.89	1.14	1.65	1.11	0.98	0.91	-0.72
Requires Improvement	1.63*	9.04	1.39*	5.76	1.18*	2.07	1.14	1.53
Inadequate	2.62*	10.08	2.39*	8.25	1.76*	4.37	1.79*	4.14
Inspection-level controls								
School % FSM	-		Y		-		Y	
Inspection type	-		Y		-		-	
Prior Ofsted rating	-		Y		-		Y	
School performance data	-		Y		-		Y	
School absences	-		Y		-		-	

(b) Predicted probabilities

	Primary				Secondary			
	M0		M1		M0		M1	
	OI	HMI	OI	HMI	OI	HMI	OI	HMI
Outstanding	7.7%	9.0%	8.0%	8.1%	10.5%	10.3%	11.0%	9.9%
Good	60.3%	47.1%	59.1%	50.9%	47.8%	42.2%	46.7%	43.4%
Requires Improvement	27.8%	35.4%	28.6%	32.8%	33.9%	35.4%	34.5%	34.7%
Inadequate	4.2%	8.6%	4.3%	8.1%	7.8%	12.1%	7.9%	12.0%
Inspection-level controls								
School % FSM	-		Y		-		Y	
Inspection type	-		Y		-		-	
Prior Ofsted rating	-		Y		-		Y	
School performance data	-		Y		-		Y	
School absences	-		Y		-		-	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level. Primary/Secondary models based upon 22,743/4,970 inspections conducted by 986/565 inspectors.

Outside home region

Table F4. The link between the inspection being conducted outside of the lead inspectors home region and school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

	Primary		Secondary	
	OR	**T-stat**	**OR**	**T-stat**
Impact of inspection outside home region (Ref: Good)				
Outstanding	0.82*	-2.20	1.40	1.93
Requires Improvement	1.06	1.02	1.22	1.91
Inadequate	1.22	1.94	0.99	-0.03
Inspection-level controls				
School % FSM	Y		Y	
Inspection type	Y		-	
Prior Ofsted rating	Y		Y	
School performance data	Y		Y	
School absences	Y		-	
Inspector gender	Y		Y	
Inspector contract	Y		Y	

(b) Predicted probabilities

	Primary		Secondary	
	Inside	**Outside**	**Inside**	**Outside**
Outstanding	8%	7%	10%	12%
Good	57%	57%	45%	41%
Requires Improvement	29%	30%	35%	38%
Inadequate	5%	6%	10%	9%
N	**15,070**	**3,175**	**3,123**	**712**

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level. Primary/Secondary models based upon 18,245/3,847 inspections conducted by 760/322 inspectors.

Table F5. The link between inspectors' phase specialism (primary / secondary) and primary school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

N = 21,437	30-69% primary		70-99% primary	
	OR	T-stat	OR	T-stat
Reference outcome = Good				
Outstanding	1.53	3.95	1.17	2.14
Requires Improvement	0.95	-0.67	0.98	-0.25
Inadequate	1.02	0.16	0.98	-0.14
Controls				
School % FSM	Yes			
Inspection type	Yes			
Prior Ofsted rating	Yes			
School performance data	Yes			
School absences	Yes			
Inspector gender	Yes			
Inspector contract status	Yes			

(b) Predicted probabilities

	Primary only	30-69% primary	70-99% primary
Outstanding	8%	10%	9%
Good	57%	56%	57%
Requires Improvement	30%	28%	29%
Inadequate	5%	5%	5%
N	**15,000**	**1,815**	**4,622**

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level.

Table F6. The link between inspectors' phase specialism (primary / secondary) and <u>secondary</u> school inspection outcomes. Multinominal regression estimates.

(a) Regression model estimates

N = 4,970	30-69% primary		70-99% primary	
	OR	T-stat	OR	T-stat
Reference outcome = Good				
Outstanding	1.02	0.12	0.85	-0.85
Requires Improvement	1.11	1.16	1.20	1.39
Inadequate	1.09	0.64	0.89	-0.47
Controls				
School % FSM	Yes			
Prior Ofsted rating	Yes			
School performance data	Yes			
Inspector gender	Yes			
Inspector contract status	Yes			

(b) Predicted probabilities

	Secondary only	30-69% primary	70-99% primary
Outstanding	11%	11%	9%
Good	46%	44%	45%
Requires Improvement	33%	35%	38%
Inadequate	10%	10%	8%
N	**2,307**	**1,953**	**710**

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level.

Team size primary

Table F7. The link between inspection team size and primary school inspection outcomes. Multinominal regression estimates.

(c) Regression model estimates

N = 21,131	2 inspectors		3 inspectors		4 inspectors		5 inspectors	
	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat
Reference outcome = Good								
Outstanding	0.90	-1.30	1.20*	2.23	1.94*	4.80	3.17*	4.50
Requires Improvement	1.11*	2.18	1.12*	2.29	1.01	0.12	1.25	1.02
Inadequate	1.81*	5.42	2.03*	6.08	1.71*	3.21	0.50	-1.28
Controls								
School % FSM	Yes							
Inspection type	Yes							
Prior Ofsted rating	Yes							
School performance data	Yes							
School absences	Yes							
Inspector gender	Yes							
Inspector contract status	Yes							

(c) Predicted probabilities

	Number of inspectors				
	1	2	3	4	5
Outstanding	8%	7%	9%	13%	17%
Good	60%	57%	55%	55%	50%
Requires Improvement	29%	30%	30%	27%	31%
Inadequate	3%	6%	6%	5%	2%
N	**5,546**	**7,184**	**7,158**	**1,093**	**150**

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level.

Table F8. The link between inspection team size and secondary school inspection outcomes. Multinominal regression estimates.

(c) Regression model estimates

	1 inspector		2 inspectors		3 inspectors		5 inspectors	
	Logit	**T-stat**	**Logit**	**T-stat**	**Logit**	**T-stat**	**Logit**	**T-stat**
Reference outcome = Good								
Outstanding	2.58*	-4.39	0.91	0.35	0.66*	2.73	1.15	-0.93
Requires Improvement	0.87	0.93	0.97	0.22	0.89	1.41	0.96	0.39
Inadequate	0.53*	3.10	0.82	1.04	0.97	0.32	0.60*	3.42
Controls								
School % FSM	Yes							
Prior Ofsted rating	Yes							
School performance data	Yes							
Inspector contract status	Yes							

(d) Predicted probabilities

	Number of inspectors				
	1	**2**	**3**	**4**	**5**
Outstanding	17%	9%	7%	9%	10%
Good	39%	42%	44%	41%	42%
Requires Improvement	35%	36%	34%	36%	38%
Inadequate	9%	12%	15%	14%	10%
N	**404**	**407**	**1,503**	**2,593**	**1,035**

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * Indicates that the estimates are statistically significant at the five percent level. Predicted probabilities generated holding other values of the covariates to their mean. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Standard errors have been clustered at the inspector level.

Appendix G. Ordinal logistic regression model estimates of differences in Overall Effectiveness judgements between OIs and HMIs

Table G1. Ordinal regression model estimates of the link between contract status and inspection outcomes. Secondary school results.

	M0		M1		M2		M3		M4		M5		M6	
	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat	OR	T-stat
HMI (ref: OI)	1.26*	3.05	1.20*	2.72	1.13	1.69	1.18*	2.14	1.18*	2.09	1.18*	2.03	1.32*	3.49
Inspection-level controls														
School % FSM	-		Y		Y		Y		Y		Y		Y	
School religion	-		Y		Y		Y		Y		Y		Y	
School gender	-		Y		Y		Y		Y		Y		Y	
Ofsted region	-		Y		Y		Y		Y		Y		Y	
Inspection type	-		-		Y		Y		Y		Y		Y	
Prior Ofsted rating	-		-		Y		Y		Y		Y		Y	
School performance data	-		-		-		Y		Y		Y		Y	
School absences	-		-		-		-		Y		Y		Y	
School % EAL	-		-		-		-		Y		Y		Y	
School % SEN	-		-		-		-		Y		Y		Y	
Inspector level controls														
Inspector gender	-		-		-		-		-		Y		Y	
Inspector phase specialism	-		-		-		-		-		-		Y	
Inspecting inside home region	-		-		-		-		-		-		Y	
Inspection experience	-		-		-		-		-		-		Y	

Notes: OR refers to the estimated odds-ratio and T-stat to the estimated t-statistics. * indicates that the estimates are statistically significant at the five percent level. Odds-ratios above one indicates that being inspected by an HMI is associated with a worse inspection outcome. Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. M0-M3 based upon 5,024 inspections conducted by 586 inspectors. M4-M6 based upon 4,899 inspections conducted by 564 inspectors. Standard errors have been clustered at the inspector level.

Appendix H. Alternative estimates for the link between lead inspector gender and short inspection outcomes for primary schools

	Watchsted sample			Extended sample		
	N	Odds ratio	T-Stat	N	Odds ratio	T-Stat
M0	8302	0.84	-1.71	8860	0.81	-2.15
M1	8302	0.83	-1.87	8860	0.80	-2.40
M2	8302	0.83	-1.87	8860	0.80	-2.40
M3	8302	0.83	-1.81	8860	0.80	-2.35
M4	8302	0.83	-1.81	8860	0.80	-2.34
M5	8302	0.82	-1.94	8860	0.79	-2.52
M6	8302	0.82	-2.00	8860	0.77	-2.68
M7	8302	0.81	-2.08	8860	0.77	-2.78

Notes: See Table 7 for details of model specifications. Model M7 adds a control for academic year, in addition to the variables controlled in model M6. Based upon short inspections conducted between September 2015 and August 2019. The outcome measure is a "negative" short inspection outcome (conversion to a full inspection leading to a downgrade in the Overall Effectiveness judgement or recommendation of an S5 inspection next due to concerns). Odds ratio below one indicates male inspectors award more lenient short inspection outcomes than their female counterpart

 @cepeo_ucl

ucl.ac.uk/ioe/cepeo