School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nses20

Are some school inspectors more lenient than others?

Christian Bokhove, John Jerrim & Sam Sims

To cite this article: Christian Bokhove, John Jerrim & Sam Sims (2023): Are some school inspectors more lenient than others?, School Effectiveness and School Improvement, DOI: 10.1080/09243453.2023.2240318

To link to this article: https://doi.org/10.1080/09243453.2023.2240318

Published online: 28 Jul 2023.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT
https://doi.org/10.1080/09243453.2023.2240318

Are some school inspectors more lenient than others?

Christian Bokhove [a], John Jerrim [b] and Sam Sims [c]

[a]Southampton Education School, University of Southampton, Southampton, UK; [b]UCL Social Research Institute, University College London, London, UK; [c]UCL Centre for Education Policy and Equalising Opportunities, University College London, London, UK

ABSTRACT

School inspections are a common feature of education systems across the world. These involve trained professionals visiting schools and reaching judgements about the quality of education they provide. Yet there is currently little academic research 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 Office for Standards in Education, Children's Services and Skills (Ofsted) employees (Her Majesty's Inspectors) are found to be harsher than those who inspect schools on a freelance basis (Ofsted Inspectors).

ARTICLE HISTORY
Received 1 March 2023
Accepted 18 July 2023

KEYWORDS
Ofsted; school inspection; consistency; accountability

Introduction

School inspections involve a team of trained inspectors visiting schools and judging the quality of education that they provide. The outcomes are often high stakes for schools and their staff (Kemethofer et al., 2017), with judgements widely reported by local media. In the extreme, inspection judgements can 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 stakeholders, including parents when choosing schools (Bokhove et al., 2023). Given the importance attached to inspection outcomes, it is vital they are as valid, consistent, and reliable as possible. Inspectorates – such as the Office for Standards in Education, Children's Services and Skills (Ofsted) in England – therefore devote significant time and resource into developing inspection frameworks, and training inspectors in their use (Ofsted, 2022).

Despite these efforts, some have questioned the validity of Ofsted inspections (whether they accurately capture school quality) and the consistency of outcomes

CONTACT John Jerrim j.jerrim@ucl.ac.uk
Author list alphabetical. All joint first authors.

Supplemental data for this article can be accessed at https://doi.org/10.1080/09243453.2023.2240318

across different inspectors (whether the same judgements would be made if the inspection were conducted by different inspectors or on different days). Despite the effort of inspectorates to develop frameworks and provide training in their use, evidence from the public administration literature questions 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). This is recognised within Ofsted's inspection handbook, which states that inspectors should draw on all the evidence they have gathered and use their professional judgement (Ofsted, 2022). Yet this has led some to question the usefulness of school inspections as a mechanism for monitoring school standards and as a force for improvement (National Education Union, 2023). 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 are assigned.

Relatively few studies have been conducted into the consistency of school inspections. In the late 1990s, 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 a strong correlation between the judgements (Pearson correlation = 0.81) but only moderate inter-rater reliability (Cohen's kappa = 0.53). They hence concluded "that OFSTED's Framework and related advice provide an effective means by which such inspectors can judge teaching with considerable reliability" (Matthews et al., 1998, p. 186). More recently, Ofsted (2017) investigated the consistency of 24 short inspections. Each school was assigned two inspectors, with their judgements compared. The same inspection outcome was reached for 22 out of the 24 schools. This research has, however, been criticised on methodological grounds (Pearson, 2018).

Other research by Ofsted has focused on the inter-rater reliability of specific inspection tasks. One example is where nine of Her Majesty's Inspectors (HMIs) undertook "workbook scrutiny", with the same documents evaluated by two or three independent inspectors (Ofsted, 2019b). "Moderate" levels of inter-rater reliability were found (Cohen's kappa around 0.5). There have been similar investigations into lesson observations (Ofsted, 2019a), with moderate-to-substantial levels of reliability found for schools (kappa around 0.6), but lower for colleges (kappa around 0.3). The same study reported greater levels of consistency between two of Her Majesty's Inspectors (HMIs) than between an HMI and a (freelance) "Ofsted inspector".

Yet there remain notable limitations with the existing evidence base. The important work previously published by Ofsted is based on a small number of inspectors, and it is not clear to what extent the results generalise across the inspection workforce. In particular, existing work has focused on HMIs – permanent Ofsted employees – with less focus on the freelancers (Ofsted Inspectors) that the inspectorate employs. Moreover, HMIs that participated in previous work would have known they were involved in a research study, potentially influencing their behaviour. Less attention has been paid to variation in inspection outcomes across different inspectors when they are conducted in a more "natural" setting (i.e., using data from live inspections). More generally, there is little existing work exploring how inspection outcomes vary depending on

the inspector(s) that schools are assigned. Finally, much work has been conducted by inspectorates themselves, rather than by independent academics.

This paper begins 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 a set of observable inspector characteristics.

Research questions

We start by exploring differences according to gender. A wide body of evidence suggests 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, p. II). Research in 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 et al., 2006). Yet there is currently no analogous evidence with respect to gender differences in the judgements made by school inspectors. Our first research question is therefore:

(RQ1) Do female inspectors make harsher or more lenient judgements about schools than their male counterparts?

Inspection outcomes may also differ between inspectors with different employment relationships with Ofsted. Broadly speaking, inspectors work for Ofsted on one of two bases. The first group are Her Majesty's Inspectors (HMIs) – civil servants who are permanent Ofsted employees and work for Ofsted as their only job. The second group are Ofsted Inspectors (OIs) who conduct inspections for Ofsted on a freelance basis. This group typically holds other jobs in the education sector (e.g., as headteachers).[1] Given their different employment circumstances, HMIs and OIs may differ in their views of what constitutes good practice, and in their understanding of young people's educational and pastoral needs. For instance, 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 could soon be) subject to Ofsted inspections themselves (in their roles elsewhere in the education sector). 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). It is plausible that such factors may influence how OIs and HMIs go about their job, leading to a difference in the judgements they reach. Research Question 2 is therefore:

(RQ2) Do 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 productivity (Burroughs et al., 2019). Moreover, employees new to their roles tend to be less confident and more liable to make mistakes than senior staff (Grohnert et al., 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 third question is therefore:

(RQ3) How are inspection outcomes linked to inspection experience of lead inspectors?

We then consider where the inspection is taking place. Ofsted's regional operating model means that inspectors 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 possible that, when an inspector works outside their home region, they come across certain practices and approaches they are not used to. There may also be regional differences in how schools operate that impact the judgements inspectors reach. We investigate this in Research Question 4:

(RQ4) 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 is whether they have a background in primary or secondary education (and thus primary or secondary inspections). Yet England has many more primary than secondary schools. This inevitably means that some inspectors who have knowledge and inspection experience in one school phase (e.g., secondary) will sometimes lead inspections in another (e.g., primary). This could impact the judgements made. For instance, those with a specialism in secondary inspections may "play it safe" when asked to inspect a primary school, given they have less 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.[2] Inspectors who usually inspect secondary schools may hence also award lower grades to primary schools. This is investigated in our fifth research question:

(RQ5) 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 1 provides 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, p. 158). This is potentially due to the benefits of pooling information, discussing the evidence, overcoming unconscious biases, and drawing on the wisdom of groups (Bang & Frith, 2017). Indeed, within the broader literature on inspection, research has found that "groups of inspectors produced more reliable assessments than individual inspectors" in the context of hospitals (Boyd et al., 2017, p. 36). Yet, in terms of optimal team size, the evidence remains inconclusive – although somewhere between five 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 on only information that is available to all

Table 1. Descriptive statistics for the distribution of inspector characteristics.

	Primary		Secondary	
	Short	Not short	Short	Not short
Lead inspector contract				
Her Majesty's Inspector	59%	20%	60%	45%
Ofsted Inspector	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				
M	33	29	19	19
SD	30	28	17	20
minimum	1	1	1	1
25^{th} percentile	11	8	6	5
50^{th} percentile	25	20	14	12
75^{th} percentile	43	42	26	26
maximum	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	**8,329**	**21,521**	**1,199**	**4,747**

inspectors (shared information bias) or to individuals "free-riding" on the effort of others (Bang & Frith, 2017). Again, we know of little analogous evidence in the context of school inspections. We thus conclude by asking:

(RQ6) Do school inspection outcomes vary by inspection team size? Do outcomes differ between teams versus individual inspections?

Data

Our data are mainly drawn from the "Watchsted" database.[3] For each inspector, this includes details of all inspections they have conducted since September 2011, drawing on the lead inspector named in the published Ofsted reports. All secondary inspections and all primary inspections done by inspectors who have conducted at least five between September 2011 and August 2019 have been extracted. When cleaning these

data, we merge together instances where a similar name is used (e.g., Ash Rahman and Ashfaq Rahman have been combined into a single record).[4]

These data have then been merged with publicly available information on inspection outcomes.[5] This was done in three steps. First, for each inspection extracted from the Watchsted database, we take the start date and restrict the data on inspection outcomes to only those inspections conducted on that date (i.e., we force an exact match on inspection start date). Second, within this subset, we fuzzy match across the databases on school name. Finally, we check that the information on inspection outcomes – including sub-judgements – is consistent. Cases were dropped in the few instances where differences were found. This process was conducted separately for primary and secondary schools. The final data set 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 (see Appendix 1 in the online supplementary material). Appendix 2 discusses our data in further detail and also provides alternative estimates using a slightly larger sample (40,959 inspections – 93% of the total). This leads to little change to our substantive results.

Appendix 3 provides details about how we have checked the quality of our data. In brief, we 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). This information was then cross-referenced against what is recorded in our data set. The level of agreement is high, with the name of the lead inspector matching on more than 97% of occasions (confidence interval 94%–100%). This provides reassurance that measurement error in our data is likely to be low.

These data were then subsequently linked to the Department for Education's (DfE) school performance tables. This includes background characteristics (e.g., admissions policy, religious denomination, school type), composition of the student body (e.g., percent of pupils eligible for free school meals (FSM), percent of pupils with English as an additional language), and national examination performance.

Our primary analysis is concerned with how the characteristics of inspectors are linked to the Overall Effectiveness judgements they make, using Ofsted's 4-point scale:

(1) Outstanding
(2) Good
(3) Requires Improvement/Satisfactory[6]
(4) Inadequate

This measure is only available for "full" inspections. Yet, since 2015, around half of all Ofsted inspections are short ("Section 8") inspections. We hence also investigate how lead inspector characteristics are linked to short inspection outcomes between January 2018 and August 2019.[7] Specifically, we create a binary measure coded 1 if the inspector decided the school should either receive a full inspection next due to concerns or immediately converted it to a full inspection with a subsequent downgrade, and zero otherwise.

The following information has been derived about individual inspectors:

- Whether an inspector is an HMI. For each inspector 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 package (Perez, 2016) was used to predict the gender of each inspector, based on first name. A small amount of manual coding has also been conducted, where results were ambiguous.
- Primary/secondary specialism. Some inspectors conduct inspections in a single school phase (primary or secondary), while others work across both. We derive a variable, based on each inspector's inspection history, identifying whether they have conducted primary inspections only, secondary inspections only, or a mix.
- Home region. Ofsted operates a regional operating model, with each inspector sitting within a regional team. It is, however, possible for inspectors to 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 inspections.[8] A binary variable is then derived, identifying for each inspection whether the inspector was working in their home region.
- 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 in the inspection report. This information is not available from the Watchsted database; we have extracted it via our own scraping of Ofsted reports (see Appendix 2 for further details).

The distribution of these variables across all inspections included in our analysis can be found in Table 1. 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 important given that – as noted in the introduction – most previous work into 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 it. The average primary inspection is conducted by someone who has led around 30 inspections since 2011, though there is quite a lot of variability around this figure (*SD* ~25). For secondary inspections, the average amount of prior lead experience is somewhat lower (an average of 19 prior inspections led). Finally, primary inspections are conducted by smaller 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 partly reflects differences in school size.

Methodology

To begin, we present cross-tabulations describing how each inspector characteristic is related to inspection outcomes. Of course, these unconditional relationships may be confounded by other factors. For instance, Ofsted could be more likely to assign inspectors with certain characteristics to inspect certain types of school.

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

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

where O_{ij} is the Overall Effectiveness judgement made by the inspector, I_j is the characteristic of the lead inspector under investigation, and X_i is 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 is other characteristics of the lead inspector (other than the characteristic under investigation), i is inspection i, j is inspector j, and k is a specific category on Ofsted's 4-point overall effectiveness scale.

The parameter of interest is β. This captures the strength of the association between the characteristic under investigation (e.g., gender) and inspection outcomes. Estimates will be presented as odds ratios, capturing the increase in the odds of receiving a *worse* inspection rating. For instance, an odds ratio of 2 will indicate that the odds of receiving an Outstanding versus a Good/RI/Inadequate rating are twice as large, conditional on the factors controlled in the model. A separate model is estimated for each characteristic under investigation. Analogous models to those presented in Equation 1 are estimated via binary logistic regression for short inspection outcomes.

The headline results reported in the main body of the paper include controls for:

- Percent of pupils eligible for FSM
- School religion
- School gender
- Ofsted region
- Inspection type
- Prior Ofsted rating
- School performance data (e.g., prior Key Stage 2 scores for primary and prior Key Stage 4 scores for secondary)
- School absences
- Percent of pupils at the school with special educational needs
- Percent of pupils who speak English as an additional language
- Other background inspector characteristics

Appendices 7, 8, and 9 provide additional results for each characteristic to illustrate the robustness of findings to several different model specifications.

To account for the nested structure of the data, standard errors are clustered at the inspector *j* level. In Appendix 4, we compare estimates to those from multilevel (random effects) models and find little substantive difference.[9] Appendix 5 presents a selection of subgroup estimates by gender and contract status. Likewise, in Appendix 6 we present alternative estimates based upon multinomial (rather than ordinal) logistic regression to investigate the sensitivity of our findings to relaxing the proportional odds assumption.

Joint effect – looking at the impact of multiple characteristics together

To investigate the combined effect of multiple inspector characteristics, we estimate an ordinal logistic regression model including the two inspector characteristics that we have found to be clearly associated with inspection outcomes (gender, HMI/OI) along with inspection team size and school/inspection controls. We then consider differences in the predicted Overall Effectiveness distribution between two hypothetical inspectors:

- Inspector A. A female HMI who is accompanied by one other inspector.
- Inspector B. A male OI who is conducting the inspection alone.

Our focus here will be primary school inspections, given the much larger sample available. 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 are undergoing the same type of inspection.

Results

RQ1: Do female inspectors make harsher judgements than their male counterparts?

Table 2 presents cross-tabulations between Overall Effectiveness grades and lead inspector characteristics. Panel (a) refers to gender.

Starting with primary schools, evidence emerges of a modest gender difference. Female lead inspectors make slightly harsher judgements about primary schools than males. For instance, male lead inspectors judged 33.1% of primary schools as Require Improvement or Inadequate, compared to 36.4% of female leads. The difference in the Inadequate grade (5.9% versus 4.5%) is notable given the relative size of the gender difference and the high-stakes consequences attached. Male lead inspectors are, on the other hand, almost 3 percentage points more likely to judge schools to be Good than females. Yet, there is little evidence of a gender gap for the Outstanding grade. Nevertheless, Table 2, Panel (a) suggests primary inspection outcomes may differ slightly by the gender of the lead inspector.

The results for secondary schools – presented on the right-hand side of Table 2, Panel (a) – are more ambiguous. 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

Table 2. Cross-tabulations between characteristics of the lead inspector and Overall Effectiveness judgements.

(a) Gender

	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
n (inspections)	**11,056**	**11,698**		**2,188**	**2,813**	

(b) Contract status

	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
n (inspections)	**17,622**	**5,139**		**2,654**	**2,370**	

(c) Inspection outside of home region

	Primary			Secondary		
	No	Yes	Difference	No	Yes	Difference
Outstanding	8.4	7.4	−1.0	9.1	14.7	5.6
Good	56.8	58.9	2.1	44.6	44.5	−0.1
Requires Improvement	29.5	28.5	−1.0	35.7	33.6	−2.1
Inadequate	5.2	5.2	0.0	10.6	7.2	−3.4
n (inspections)	**15,925**	**3,347**		**3,161**	**735**	

(d) Primary/secondary specialism

	Primary			Secondary		
	Split	Primary only		Split	Second only	
Outstanding	11.0	7.2	−3.8	10.2	11.1	0.9
Good	52.9	58.2	5.3	43.9	46.1	2.2
Requires Improvement	30.1	29.5	−0.6	35.3	33.1	−2.2
Inadequate	6.0	5.0	−1.0	10.6	9.8	−0.8
n (inspections)	**1,912**	**15,871**		**1,976**	**2,308**	

Note: OI = Ofsted inspector; HMI = Her Majesty's Inspector.

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 quite small.

To what extent might these unconditional results be driven by selection? Are the harsher judgements made by female inspectors due to them being assigned more challenging primary schools to inspect? Two pieces of evidence are presented. First, Table 3 compares the distribution of observable characteristics of the primary schools inspected by male and female lead inspectors. If the gender difference in Overall Effectiveness grades for primary schools observed in Table 2, Panel (a) is due to selection effects, one would expect to see female inspectors being disproportionately assigned to inspect lower "quality" schools (e.g., those with lower prior inspection ratings, worse performance in national examinations, higher absence levels). Table 3 provides little indication that this is the case; the distribution of inspection is similar across male and female lead inspectors.

Table 3. Differences in inspection assignments by gender and contract status of the lead inspector (primary schools).

	Gender		Her Majesty's Inspector	
	Female	Male	No	Yes
Inspection type				
Section 5	68%	70%	74%	49%
Requires Improvement reinspection	20%	18%	18%	22%
Academy first Section 5	5%	4%	3%	10%
Section 8 deemed Section 5	4%	4%	4%	7%
Serious weakness inspection	1%	1%	1%	4%
Exempt school inspection	2%	2%	0%	8%
Section 8 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%
Requires Improvement	43%	45%	47%	32%
Inadequate	4%	4%	2%	14%
Missing	3%	3%	3%	5%
Free school meals (FSM) quintile				
Quintile 1 (Low FSM)	16%	17%	17%	15%
Quintile 2	19%	20%	20%	18%
Quintile 3	21%	21%	21%	20%
Quintile 4	22%	22%	21%	24%
Quintile 5 (High FSM)	22%	20%	20%	23%
Missing	0%	0%	0%	0%
School absence quintile				
Quintile 1 (low absences)	20%	21%	21%	19%
Quintile 2	23%	22%	23%	21%
Quintile 3	23%	23%	23%	24%
Quintile 4	21%	21%	21%	22%
Quintile 5 (high absences)	13%	13%	13%	14%
Missing	0%	0%	0%	0%
Key Stage 2 English quintile				
Quintile 1 (low achievement)	24%	24%	23%	31%
Quintile 2	21%	20%	21%	20%
Quintile 3	17%	17%	18%	14%
Quintile 4	17%	16%	17%	14%
Quintile 5 (high achievement)	12%	13%	13%	9%
Missing	9%	10%	10%	10%
Key Stage 2 maths quintile				
Quintile 1 (low achievement)	24%	23%	22%	31%
Quintile 2	19%	20%	20%	18%
Quintile 3	20%	18%	19%	20%
Quintile 4	15%	15%	16%	12%
Quintile 5 (high achievement)	12%	13%	13%	10%
Missing	9%	10%	10%	10%

Second, Table 4 presents estimates from our ordinal regression models. The top row refers to those for gender, with values below 1 indicating that female lead inspectors make harsher judgements than their male counterparts (conditional on the controls).

The estimated odds ratio is 0.84 and is statistically significant at the 5% level. This confirms that female lead inspectors tend to award lower inspection grades to primary schools than male inspectors, even after controlling for observable differences in the schools they are assigned to inspect (and other inspector-level characteristics).[10] Moreover, in Appendix 9 (Table 9.3), we illustrate how the odds ratio is very stable across multiple specifications – suggesting that any unobserved confounding would have to be generated by a factor that is strongly associated with inspection outcomes, but orthogonal to a school's intake, performance in examinations, pupil absences, and previous

Table 4. Ordinal logistic regression model results of the association between lead inspector characteristics and Overall Effectiveness grades.

	Primary		Secondary	
	OR	*t*	*OR*	*t*
Gender (Ref: Male)				
Female	0.84*	−3.18	1.09	1.16
Contract status (Ref: Ofsted Inspector)				
Her Majesty's Inspector	1.45*	6.21	1.32*	3.49
Experience (Ref: Bottom quintile)				
Quintile 2	0.94	−0.6	0.72	−1.95
Quintile 3	1.00	0.03	0.71*	−2.13
Quintile 4	1.07	0.69	0.82	−1.08
Quintile 5 (most experienced)	0.93	−0.64	0.89	−0.53
Inspection outside home region (Ref: No)				
Yes	1.13*	2.50	0.99	−0.14
Phase specialism (ref: primary/secondary only)				
30%–69% primary	0.86	−1.91	0.92	−1.06
70%–99% primary	0.95	−0.83	1.03	0.21
Team size				
1 inspector	Reference		0.43*	−4.92
2 inspectors	1.25*	5.29	0.90	−0.93
3 inspectors	1.26*	5.23	1.08	0.93
4 inspectors	1.05	0.56	Reference	
5 + inspectors	0.68	−1.60	0.83*	−2.10

Note: Data based upon inspections conducted between the 2011/12 to 2018/19 academic years. Estimates based on 22,754 inspections conducted by 983 inspectors. Standard errors have been clustered at the inspector level. Models control for percent of pupils eligible for free school meals, school religion, school gender, region, inspection type, prior Ofsted rating, school performance data, school absences, percent of pupils at the school with special educational needs, percent of pupils who speak English as an additional language, and other inspector characteristics.

*$p < .05$.

Table 5. Logistic regression model estimates of the association between lead inspector characteristics and primary school short inspection outcomes.

	n	*OR*	*t*
Gender (Ref: Male)			
Female	3,605	0.75*	2.17
Contract (Ref: Ofsted inspector)			
Her Majesty's Inspector	3,627	1.63*	3.56
Experience (Ref: Bottom quintile)			
Quintile 2	3,627	0.65	−1.85
Quintile 3		0.84	−0.73
Quintile 4		1.05	0.22
Quintile 5 (most experienced)		1.22	0.92
Inspection outside home region (Ref: No)			
Yes	3,311	1.41	1.08
Phase specialism (ref: primary only)			
30%–69% primary	3,594	1.41	1.64
70%–99% primary		1.36	1.77
Team size (Ref: 1 inspector)			
2 inspectors	3,511	0.99	−0.05

Note: Sample restricted to short primary school inspections conducted between January 2018 and August 2019. Dependent variable coded 1 if the short inspection resulted in the outcome "Section 5 next due to concerns" or was immediately converted to a full inspection resulting in a judgement of Inadequate or Requires Improvement, and zero otherwise. Models have been estimated separately for each characteristic. Models control for percent of pupils eligible for free school meals, school religion, school gender, region, inspection type, prior Ofsted rating, school performance data, school absences, percent of pupils at the school with special educational needs, percent of pupils who speak English as an additional language, and other inspector characteristics.

*$p < .05$.

Ofsted grades. It is not clear what such a characteristic could be. Our interpretation is hence that the gender difference we observe in primary inspection outcomes is unlikely to be driven by inspector selection.

Table 5 turns to analogous modelling results for short primary school inspections, with the top row providing those for gender. Odds ratios below 1 indicate that short inspections led by females are more likely to result in a negative outcome (i.e., a full Section 5 inspection within the next year due to concerns) than those led by males.

Evidence again emerges that female lead inspectors reach slightly harsher verdicts than their male counterparts. The odds ratio sits around 0.75, indicating that the odds of a negative outcome from a short primary inspection are around 25% lower for males than females. In absolute terms, this represents a modest difference of between 2 and 3 percentage points (9.6% of short inspections with a male lead led to a negative outcome for the school compared to 12.1% of those with a female lead).

There is no evidence of such a gender difference with respect to secondary inspections. The ordinal logistic regression model presented in Table 4 illustrates that, for secondary schools, the odds ratio stands at 1.09 and is not statistically significant at conventional thresholds. Additional analyses we have performed for short secondary inspections have also proven inconclusive. Hence, our finding of a small gender difference in inspection outcomes seems to hold only for primary schools.

RQ2: Do Ofsted inspection judgements differ between OIs and HMIs?

Table 2, Panel (b) illustrates the distribution of inspection outcomes by contract status (HMI versus OI). Starting with primary schools, HMIs are around 13 percentage points less likely to judge a school to be Good than OIs (60% versus 47%). They are much more likely to award Requires Improvement (35% versus 28%) and Inadequate (9% versus 4%) grades instead. For secondary schools, HMIs judge fewer to be Good than OIs (42% versus 48%) and place more in the Inadequate category (12% versus 8%). Nevertheless, the difference between HMI and OI lead inspectors is greater for primary than secondary schools.

The right-hand columns in Table 3 provide one possible explanation of this result – Ofsted deploys OIs and HMIs to different inspection tasks. In particular, HMIs are more likely to be assigned to schools with lower performance in national examinations and those that were judged to be Inadequate during their last inspection. Thus, the "contract status" row in Table 4 illustrates whether we continue to find a difference in Overall Effectiveness outcomes between HMI and OI after we have controlled – as far as possible – for differences in their inspection tasks. Odds ratios greater than 1 indicate that HMIs tend to provide harsher inspection judgements than OIs.

There are two key points. First, the relationship between contract status and inspection outcomes is strong and statistically significant. Roughly speaking, the odds of a primary school being placed in a lower Overall Effectiveness category are around 50% higher if the lead inspector is an HMI rather than an OI. Second, as Appendix 9 illustrates (see Table 9.7), the inclusion of inspection, school, and inspector controls only slightly weakens the relationship across model specifications with the estimated odds ratio consistently between 1.4 and 1.5). This suggests the result is not being solely driven by the selection of HMIs/OIs into different types of inspection, at least in terms 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 factors we cannot observe (and is not well proxied by our controls). Our analysis of short primary inspections in Table 5 produces similar results; the odds of a negative outcome are around 1.6 times higher if conducted by an HMI rather than an OI (conditional on the controls). In absolute terms, this suggests that about 13% of short inspections led by HMIs result in a full inspection next due to concerns, compared to 9% of those led by OIs.

Analogous results for secondary schools in Table 4 point towards a similar – although slightly weaker – relationship ($OR = 1.32$), with the odds ratio fluctuating slightly (between 1.13 and 1.32) depending on the exact specification used (see Appendix 9, Table 9.8). Moreover, alternative estimates based upon multinomial (rather than ordinal) logistic regression in Appendix 6 make clear that for secondaries, the main point of difference between HMIs and OIs is with respect to Good and Inadequate judgements.

RQ3: How are inspection outcomes linked to inspection experience of lead inspectors?

The next set of estimates in Table 4 presents results for inspector experience. Those for both primary and secondary schools suggest there is no clear relationship with inspection outcomes. Moreover, in Appendix 9 (Table 9.12), we illustrate how this holds true regardless of the model specification used. Similar results also emerge for primary school short inspections in Table 5.

RQ4: Do inspectors judge schools more harshly when they are working outside of their home region?

Table 2, Panel (c) presents a cross-tabulation between whether the inspection was conducted inside the inspector's home region and Overall Effectiveness judgement. For primary schools, the distribution is very similar whether the inspection was conducted within inspector's home region or not. The regression model estimates presented in Table 4 suggest that, although statistically significant due to the large sample size, any association here is weak ($OR = 1.13$). For primary school short inspections (Table 5), only 96 out of the 3,311 within our database have been conducted outside of the lead inspectors home region meaning one should not read too much into these estimates (the odds ratio is quite sizeable at 1.41 but not statistically significant).

For secondary schools, the cross-tabulations in Table 2, Panel (c) suggest that those conducted outside the inspectors' home region are more likely to be rated Outstanding (15% versus 9%) and less likely to be rated Inadequate (7% versus 11%). However, the estimated odds ratio quickly approaches 1 in the ordinal logistic regression model results presented in Table 4. We thus conclude that there is no evidence that the inspection judgements secondary schools receive are related to whether the lead inspector was working in their home region or not.

RQ5: Do inspectors with a specialism in secondary school inspections judge primary schools more harshly than inspectors with a primary specialism?

Table 2, Panel (d) presents a cross-tabulation between the percent of primary school inspections each inspector has conducted during their career and Overall Effectiveness judgements. No clear relationship is found for either primary or secondary schools. This continues to hold after controlling for a set of school, inspection, and inspector characteristics within our ordinal regression models in Table 4. For short primary school inspections (Table 5), there is some suggestion of a difference between those who have only conducted primary inspections and those who have conducted a mix of primary and secondary inspections (odds ratio around 1.4), though these differences are only statistically significant at the 10% level. Overall, it seems there is relatively little evidence of an association between whether inspectors have a specialism in the primary/secondary sector and inspection outcomes.

RQ6: Do school inspection outcomes vary by inspection team size? Do outcomes differ between team versus individual inspections?

Table 6 presents a cross-tabulation between inspection team size and Overall Effectiveness grades. For primary inspections, larger teams are less likely to reach a Good judgement and are more likely to rate schools as Inadequate or Requires Improvement. The difference between a single inspector and two to three inspectors remains statistically significant in our ordinal regression models, with the estimated odds ratio standing around 1.25. Appendix 9 illustrates that this result is robust to a wide set of alternative model specifications (see Table 9.23). Additional multinomial logistic regression estimates (see Appendix 6, Table 6.7) 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

Table 6. 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**
(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**

Note: Figures refer to column percentages.

Table 7. Predicted distribution of primary school inspection outcomes for two hypothetical inspectors. Ordinal logistic regression model 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 Section 5 next due to concerns. (Jan18–Aug19)	15.5%	9.7%	1.60
Inspector characteristics			
Team size	2 inspectors	1 inspector	
Contract status	Her Majesty's Inspector	Ofsted Inspector	
Gender	Female	Male	

Note: Model controls for percent of pupils eligible for free school meals, 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, whether the inspection was conducted after 2018, school religion, school gender composition, Key Stage 1 scores, and percent of pupils with special educational needs.

around 6% when it is conducted by a team of two, three, or four inspectors. Interestingly, however, we find no association between short primary inspection outcomes and team size.

For secondary schools, very small teams (one inspector) and large teams (five inspectors or more) seem to make slightly less harsh judgements than secondary inspections conducted by a team of four. The ordinal regression estimates in Table 4 are consistently statistically significant at the 5% level, with the estimated odds ratio around 0.4 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 is independently associated with Overall Effectiveness judgements, over and above our school- and inspection-level controls.

Joint effect – looking at the impact of multiple characteristics together

To conclude, we examine the combined effect of multiple inspector (and inspection team) characteristics at the same time. Results can be found in Table 7. This part of our analysis focuses on primary schools only, given this is where we have found the most convincing evidence of difference in preceding subsections.

There is a clear, sizeable difference in inspection outcomes reached by our two hypothetical lead inspectors. Inspector A is around twice likely to award an Inadequate judgement than Inspector B (9.0% versus 4.5%), while being around half as likely to judge a primary school to be Outstanding (4.5% versus 9.1%). Likewise, almost half of the primary schools inspected by Inspector A will be judged to be Inadequate or Requires Improvement, compared to less than a third of those inspected by Inspector B. The analogous difference for short primary inspection outcomes – with respect to recommending a full Section 5 inspection next due to concerns – is 15.5% for Inspector A compared to 9.7% for Inspector B. Appendix 9 provides an alternative version of Table 7 using multinomial – rather than ordinal – logistic regression modelling (see Appendix 9, Table 9.25). The clearest point of difference when using this alternative analytic approach is an increase in the difference between the two hypothetical inspectors awarding the Inadequate grade (13.3% versus 3.4%).

Conclusions

School inspections are a common feature of education systems across the globe. Although such inspections come in different shapes and sizes, in some countries – such as England – they are a key part of the accountability system. Ofsted – the school inspectorate in England – is one example where a team of inspectors make high-stakes judgements about schools. Yet relatively few studies have been conducted into variation in school inspection outcomes, with most existing work limited in scope and conducted by school inspectorates themselves (e.g., Matthews et al., 1998; Ofsted, 2017).

This paper has sought to address this gap in the literature. Using data from more than 30,000 school inspections conducted over an 8-year period, we have produced the first evidence on how school inspection outcomes are linked to characteristics of the lead inspector. Robust evidence emerges that male inspectors make more lenient judgements about primary schools than females. Although the magnitude of the gender differences is relatively small, it is most apparent at the high-stakes (Inadequate) grade. Much larger differences are observed between inspectors working under different contractual arrangements (HMIs versus OIs), with the former reaching harsher judgements than the latter. Likewise, inspection team size also appears to be independently associated with Overall Effectiveness grades. In contrast, there is little – or at best mixed – association 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).

Previous research published by Ofsted has found short inspections carried out by different inspection teams usually reach the same judgement (Ofsted, 2019a), and has claimed that "Her Majesty's Inspectors (HMI) can assess the quality of education by using workbook scrutiny indicators and they do so reliably" (Ofsted, 2019b, p. 1). At first glance, this appears at odds with our findings. However, it is important to note that the research Ofsted has published into workbook scrutiny and lesson observations actually shows there to be non-trivial differences in the opinions formed by different inspectors even when they are looking at the same piece of evidence – suggesting that the headline claim that HMIs can do such tasks "reliably" may be somewhat oversold. Moreover, the previous work conducted by Ofsted has only utilised a small number of HMIs. Our findings – particularly the sizeable gap in judgements made by HMIs and OIs – suggest that Ofsted's previous work may lack external validity; that results from their studies cannot necessarily be generalised to the inspection workforce as a whole. Indeed, it is vital that any future research conducted into school inspection consistency and reliability involves a truly representative cross-section of inspectors, rather than being restricted to a small number of selected individuals. This in turn motivates the need for future studies involving the different inspection teams making independent judgements about the same school on different days, and further analysis similar to ours that monitors what is happening across a wide array of live inspections on the ground.

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 (Bokhove et al., 2023), while women have higher levels of conscientiousness (Verbree et al., 2023). In job promotions, Hartman et al. (1991) argued that it is "predominantly the gender stereotype of the ratee's personal characteristics rather than the ratee's gender that influences the promotion process" (p. 285). 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. In a similar vein, Friesdorf et al. (2015) note how men have a stronger preference for utilitarian judgements (i.e., consider the overall consequences of an action) over deontological judgements (i.e., consider the actions consistent with moral norms) compared with women when faced with moral dilemmas. 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 we do 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.

These findings should 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 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 during our analyses and have attempted to control for such differences via estimation of regression models. Nevertheless, we recognise this may only partially overcome such issues. Second, we have only considered variation by a limited set of observable inspector characteristics. Arguably, there are likely to be more important sources of variation in inspection outcomes in terms of things we cannot observe, such as inspectors' personalities and professional history (e.g., whether the inspector has previously led a challenging school). This should be a key line of inquiry in future research. Third, it has not been possible with the data currently available to understand what may be driving between-inspector variation in school inspection outcomes. Future research – both quantitative and qualitative – should seek to better understand the mechanisms behind the differences that we observe. Finally, a new inspection framework was introduced by Ofsted in September 2019, which puts less emphasis on performance in national examinations and more on the quality of the curriculum. Unfortunately, only 6 months of inspection data are available from this new framework before the COVID-19 pandemic disrupted school inspections for the following 2 years. Our analysis

has thus been restricted to before the most recent framework change. However, given that our analytic sample covers an 8-year period during which multiple changes were made to how school inspections were conducted (including previous framework changes), we do not believe different findings would emerge now. Nevertheless, once data from further inspections are available under the new framework (outside of the pandemic era), we believe it is important that Ofsted publishes an update building on 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 recognise that inspectors should use their professional judgements when interpreting the evidence collected, with the variation we observe perhaps just 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 not an easy question to answer and is open to 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 judgement with the highest stakes, the Inadequate grade. Given the consequences of receiving an Inadequate judgement, almost any variation across inspectors in reaching this decision might be considered an issue.

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. It is only with such evidence that an open and informed debate can be had about such issues. Indeed, if governments are to have a school inspection regime with high-stakes outcomes, then it is vital that they are proven to have a high degree of validity, reliability, and consistency. Our findings suggest that, in the case of Ofsted, the high-stakes consequences attached to certain inspection outcomes may need to be adjusted downward. At the same time, it is equally important to ensure that inspections and inspectorates are appropriately resourced to deliver the level of reliability and consistency that government requires.

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 in a quicker, simpler way than is currently possible. Likewise, more needs to be documented, investigated, and discussed about inspector deployment – how exactly are inspectors assigned to different inspection tasks? Finally, Ofsted might consider publishing further details about its quality assurance processes, particularly with respect to what happens when schools receive an Inadequate grade.

Notes

1. 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 (Richardson, 2015).
2. In 2020, 88% of primary schools were rated as good or outstanding, compared to 76% of secondary schools (Ofsted, 2020).
3. Available from https://perspective.angelsolutions.co.uk/Perspective/Login.aspx?ReturnUrl=%2fPerspective%2fLiteUsers%2fOfsted%2f

4. One would ideally have access to additional information about inspectors to ensure that the merged cases refer to the same individual. Unfortunately, very little such time-invariant information about inspectors is available for us to use. As our analysis focuses on inspector characteristics – rather than individual inspectors – the impact of any incorrect merges (e.g., Ash Rahman and Ashfaq Rahman being different people) is likely to be small. In particular, our point estimates will be largely unchanged, while reported standard errors are likely to be slightly conservative (due to underestimation of the "cluster" – i.e., inspector – sample size).
5. Available from https://www.gov.uk/government/statistical-data-sets/monthly-management-information-ofsteds-school-inspections-outcomes
6. The "Satisfactory" grade was replaced with the "Requires Improvement" grade in 2012, on the basis that the original label was thought to be lacking in ambition (Ofsted, 2012).
7. Outcomes from short inspections were different between September 2015 and December 2017, when they were either immediately converted into a full inspection or the Good grade was retained.
8. 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".
9. We find that about 9.5% of the variation in Overall Effectiveness judgements occurs between inspectors for primary schools (regardless of whether controls are included in the model or not), compared to between 5% and 7% for secondary schools (depending on whether controls are included). The estimated intra-class correlation is slightly higher for short primary inspections (around 11%–12%) but lower for secondary inspections (between 0% and 5%).
10. We have also re-estimated our analytic models using multinomial (rather than ordinal) logistic regression – see Appendix 6. These confirm that there is little evidence of a gender difference when it comes to the Good/Outstanding distinction, but more so for Good/RI/Inadequate judgements. This is consistent with the descriptive results presented in Table 2, Panel (a).

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.

Disclosure 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.

Notes on contributors

Christian Bokhove is a professor in Mathematics Education at the University of Southampton. He is a specialist on international comparisons in mathematics education, the use of technology, and innovative methodologies.

John Jerrim is a professor of Education and Social Statistics at University College London. He specialises in international comparisons of educational achievement and applied quantitative education research.

Sam Sims is an assistant professor at University College London. He specialises in research about teachers and applied quantitative education research.

ORCID

Christian Bokhove http://orcid.org/0000-0002-4860-8723

References

Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, *4*(8), Article 170193. https://doi.org/10.1098/rsos.170193

Bokhove, C., Jerrim, J., & Sims, S. (2023). How useful are Ofsted inspection judgements for informing secondary school choice? *Journal of School Choice*, *17*(1), 35–61. https://doi.org/10.1080/15582159.2023.2169813

Boyd, A., Addicott, R., Robertson, R., Ross, S., & Walshe, K. (2017). Are inspectors' assessments reliable? Ratings of NHS acute hospital trust services in England. *Journal of Health Services Research & Policy*, *22*(1), 28–36. https://doi.org/10.1177/1355819616669736

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 *IEA Research for Education: Vol. 6. Teaching for excellence and equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). 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), 157–176. 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

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, *41*(5), 696–713. https://doi.org/10.1177/0146167215575731

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

Grohnert, T., Meuwissen, R. H. G., & Gijselaers, W. H. (2019). Enabling young professionals to learn from errors – The role of a supportive learning climate in crossing help network boundaries. *Vocations and Learning*, *12*(2), 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*(4), 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*(5–6), 285–300. https://doi.org/10.1007/BF00289757

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

Jerrim, J., & Sims, S. (2019). *The Teaching and Learning International Survey (TALIS) 2018*. Department for Education. 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*(4), 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. C., 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*, Article 42. https://doi.org/10.1186/1472-6920-6-42

National Education Union. (2023). *Replace Ofsted*. https://neu.org.uk/campaigns/replace-ofsted

Office for Standards in Education, Children's Services and Skills. (2012, January 16). *Ofsted scraps 'satisfactory' judgement to help improve education* [Press release]. https://www.gov.uk/government/news/ofsted-scraps-satisfactory-judgement-to-help-improve-education

Office for Standards in Education, Children's Services and Skills. (2017). *Do two inspectors inspecting the same school make consistent decisions? A study of the reliability of Ofsted's new short inspections* (Ofsted Research Report No. 170004). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/596708/Reliability_study_-_final.pdf

Office for Standards in Education, Children's Services and Skills. (2019a). *How valid and reliable is the use of lesson observation in supporting judgements on the quality of education?* (Ofsted Research Report No. 190029). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936246/Inspecting_education_quality_Lesson_observation_report.pdf

Office for Standards in Education, Children's Services and Skills. (2019b). *Workbook scrutiny: Ensuring validity and reliability in inspections* (Ofsted Research Report No. 190028). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936240/Inspecting_education_quality_workbook_scrutiny_report.pdf

Office for Standards in Education, Children's Services and Skills. (2020, October 30). Ofsted inspections illustrate high proportion of good or outstanding schools. *The Education Hub*. https://educationhub.blog.gov.uk/2020/10/30/ofsted-inspections-illustrate-high-proportion-of-good-or-outstanding-schools/

Office for Standards in Education, Children's Services and Skills. (2022). *School inspection handbook*. https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook

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*. https://www.researchgate.net/publication/327894743_A_review_of_Ofsted's_test_of_the_reliability_of_short_inspections

Perez, I. (2016). *Gender-guesser*. https://pypi.org/project/gender-guesser/

Powell, D., & Lorenz, R. (2019). The effect of team size on the performance of continuous improvement teams: Is seven really the magic number? In F. Ameri, K. E. Stecke, G. von Cieminski, & D. Kiritsis (Eds.), *IFIP advances in information and communication technology: Vol. 566. Advances in production management systems: Production management for the factory of the future* (pp. 69–76). Springer. https://doi.org/10.1007/978-3-030-30000-5_9

Richardson, H. (2015, June 19). Ofsted purges 1,200 'not good enough' inspectors. *BBC News*. https://www.bbc.co.uk/news/education-33198707

Richmond, T. (2019). *Requires Improvement. A new role for Ofsted and school inspections*. EDSK. 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, March 7). HMCI's commentary: New research into short school inspections. *GOV.UK*. 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. https://doi.org/10.5771/0935-9915-2007-3-251

Verbree, A.-R., Hornstra, L., Maas, L., & Wijngaards-de Meij, L. (2023). Conscientiousness as a predictor of the gender gap in academic achievement. *Research in Higher Education*, *64*(3), 451–472. 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